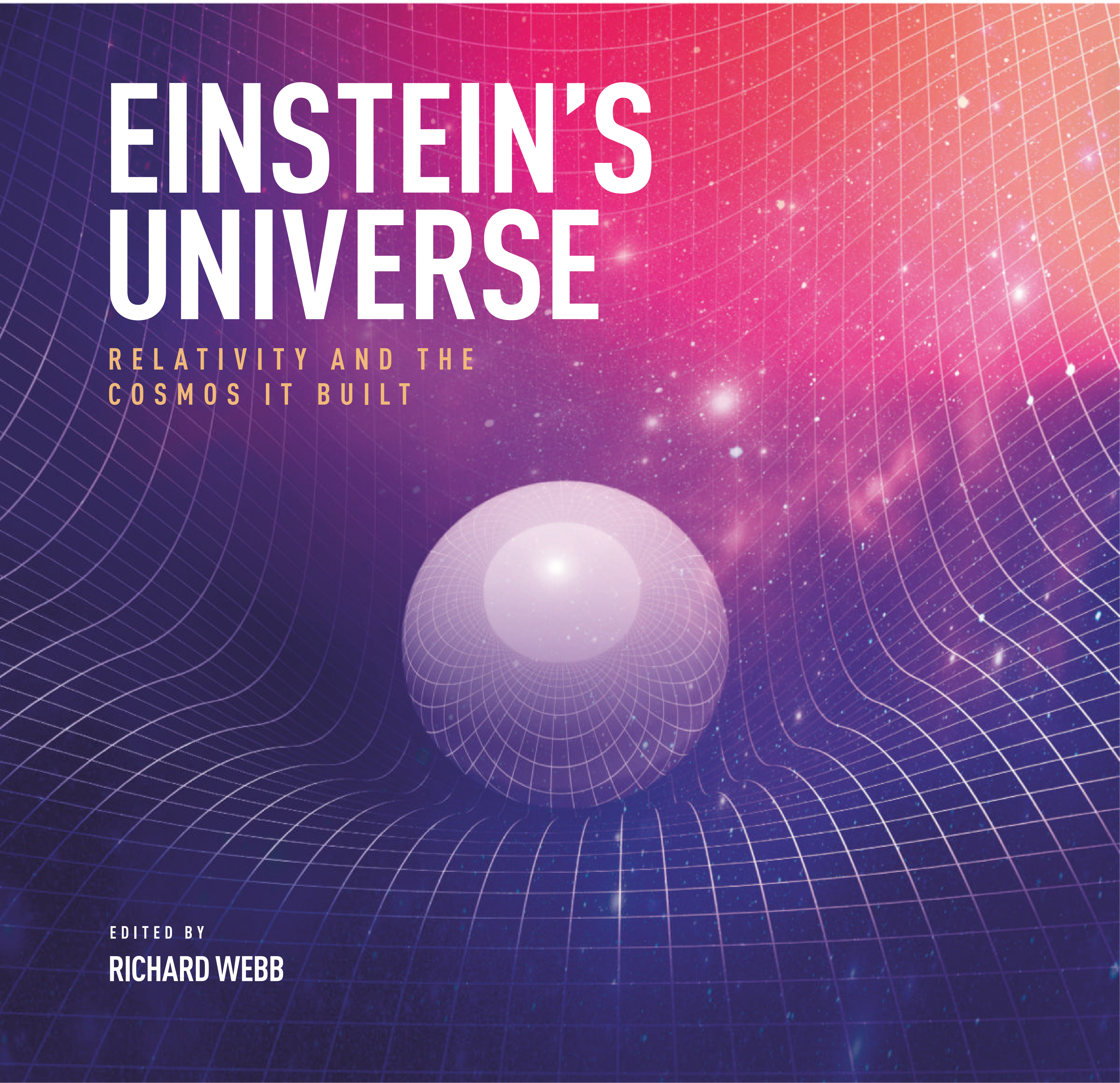
New Scientist

ESSENTIAL GUIDE Nº10

BASICS OF RELATIVITY
THE BIG BANG UNIVERSE
DARK MATTER AND DARK ENERGY
BLACK HOLES
GRAVITATIONAL WAVES
THE HUNT FOR QUANTUM GRAVITY
AND MORE



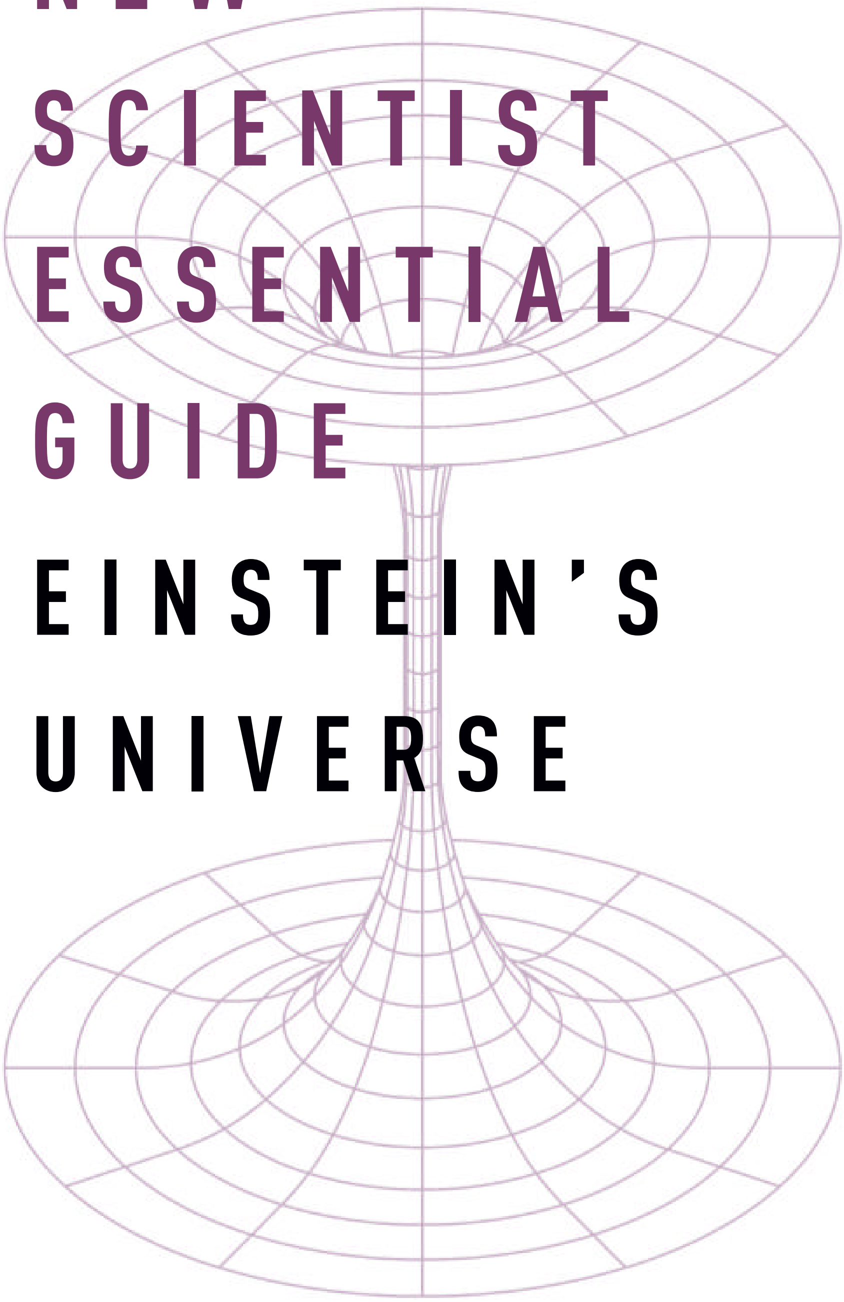
EINSTEIN'S UNIVERSE

RELATIVITY AND THE COSMOS IT BUILT

EDITED BY
RICHARD WEBB

NEW SCIENTIST ESSENTIAL GUIDE

EINSTEIN'S UNIVERSE



TO SAY Albert Einstein is a towering genius of modern physics is not to say anything anyone didn't know. But it is hard to get your head around the full scope of the cosmic revolution his theories of relativity wrought in the early years of the 20th century. It is no exaggeration to say that today we live in Einstein's universe.

This latest *New Scientist Essential Guide,* the tenth in the series, aims to give an overview of that revolution, from the basics of Einstein's special and general theories of relativity to all the strange and wondrous discoveries we have made off the back of them: the expanding big bang universe, the mysterious entities of dark matter and dark energy, black holes and the ripples in space and time known as gravitational waves.

Then, of course, there is the small matter of the deficiencies of Einstein's universe – above all, the way this epic vision fails to gel with quantum theory, our picture of material reality at its very smallest. I hope you enjoy finding out more about all these facets of Einstein's universe. Feedback is welcome at essentialguides@newscientist.com, and previous editions in the *Essential Guide* series can be bought by visiting shop.newscientist.com. *Richard Webb*

NEW SCIENTIST ESSENTIAL GUIDES
NORTHCLIFFE HOUSE, 2 DERRY STREET,
LONDON, W8 5TT
+44 (0)203 615 6500
© 2021 NEW SCIENTIST LTD, ENGLAND
NEW SCIENTIST ESSENTIAL GUIDES
ARE PUBLISHED BY NEW SCIENTIST LTD
ISSN 2634-0151
PRINTED IN THE UK BY
PRECISION COLOUR PRINTING LTD AND
DISTRIBUTED BY MARKETFORCE UK LTD
+44 (0)20 3148 3333

ABOVE: KOVALTO1/ISTOCK
COVER: BEHOLDINGEYE/ISTOCK

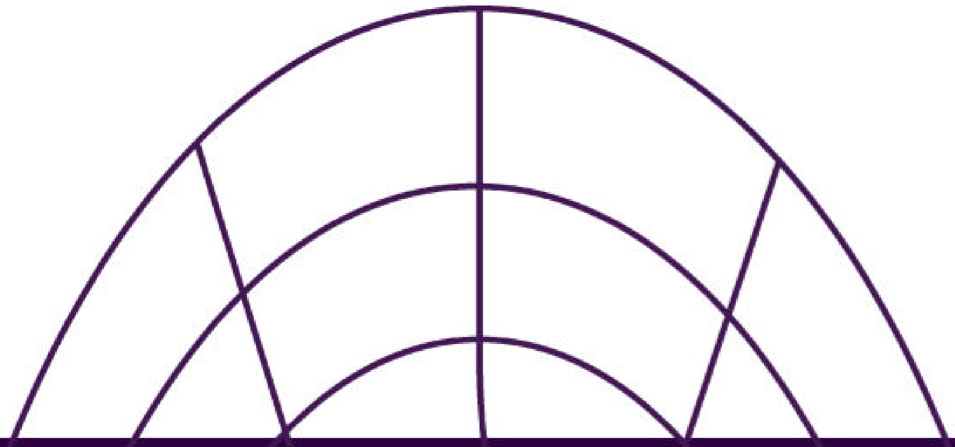
EDITOR Richard Webb
DESIGN Craig Mackie
SUBEDITOR Bethan Ackerley
PRODUCTION AND APP Joanne Keogh
TECHNICAL DEVELOPMENT (APP) Amardeep Sian
PUBLISHER Nina Wright
EDITOR-IN-CHIEF Emily Wilson
DISPLAY ADVERTISING +44 (0)203 615 6456
displayads@newscientist.com

ABOUT THE EDITOR

Richard Webb is executive editor of *New Scientist*

ADDITIONAL CONTRIBUTORS

Anil Ananthaswamy, Jacob Aron, Stephen Battersby, Michael Brooks, Mark Buchanan, Marcus Chown, Stuart Clark, Leah Crane, Daniel Cossins, Pedro Ferreira, Amanda Gefter, John Gribbin, Lisa Grossman, Stephen Hawking, Anna Ijjas, Valerie Jamieson, Eugene Lim, Michael Marshall, Joshua Sokol, Richard Webb, Chelsea Whyte



CHAPTER 1

BASICS OF RELATIVITY

Albert Einstein's towering reputation in modern physics rests above all on a theory he presented to the world in 1915. Explaining the force of gravity through matter warping space-time, general relativity revolutionised our understanding of the cosmos – but Einstein was only building on work begun 10 years earlier.

p. 6 The special theory of relativity
p. 9 Albert Einstein: A life
p. 10 What is space-time?
p. 11 The general theory of relativity
p. 12 Eddington's experiment
p. 14 Newton vs Einstein
p. 16 Einstein's universe: A timeline

CHAPTER 2

THE BIG BANG UNIVERSE

The most striking implications of general relativity only started to become clear in the 1920s, with observations indicating that the universe was expanding. Einstein himself was initially sceptical, and it wasn't until a discovery in the 1960s that the idea of a cosmos that began in a big bang became widely accepted. That raised the question of what this all meant.

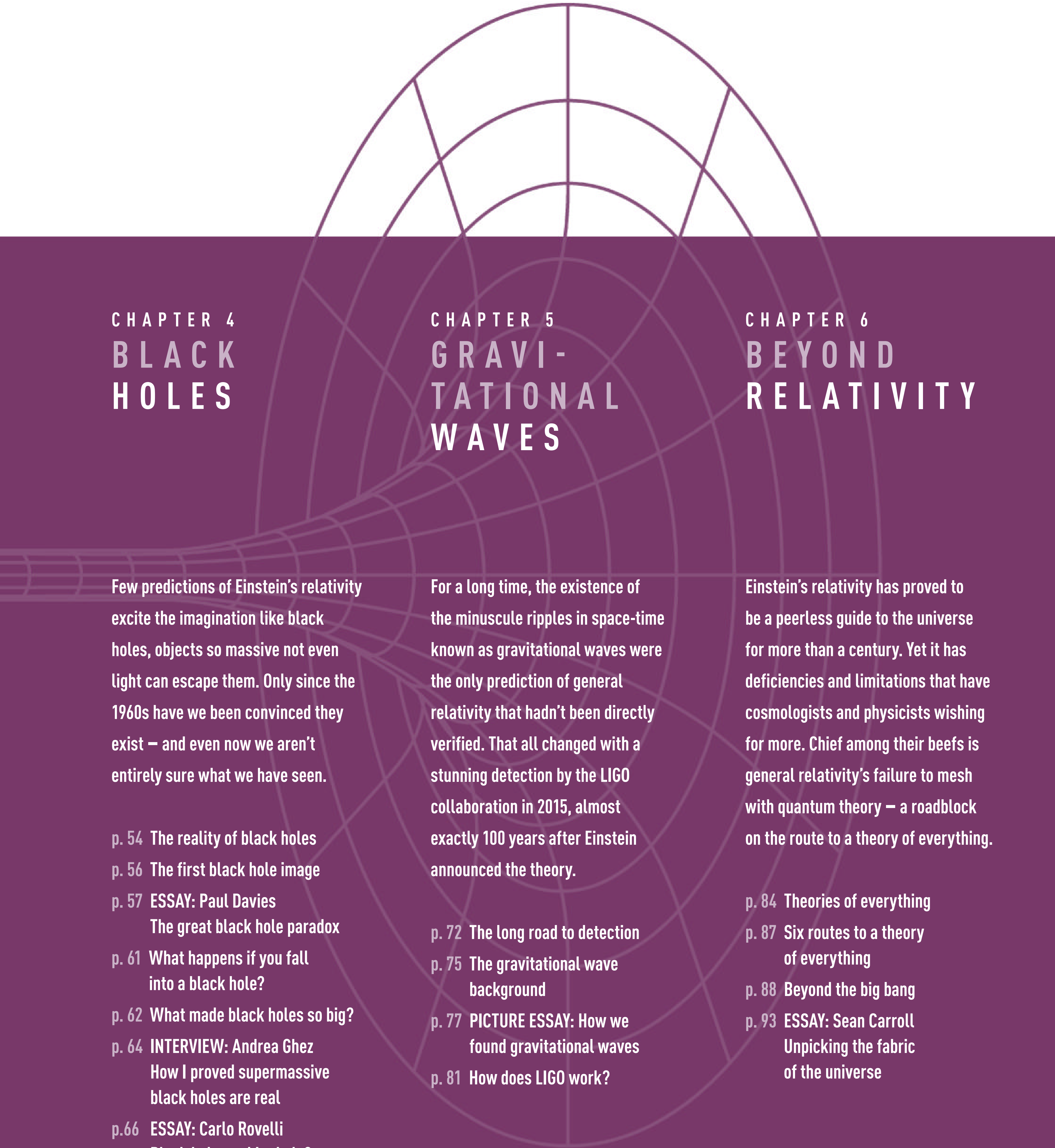
p. 20 The expanding universe
p. 23 ESSAY: Jim Peebles
How the big bang was born
p. 26 The cosmic microwave background
p. 28 What exactly was the big bang?
p. 30 Inflation and the multiverse

CHAPTER 3

DARK MATTER AND DARK ENERGY

Our standard model of cosmology based on general relativity works very well – if you accept that 95 per cent of the universe comes in two unknown forms, dark matter and dark energy, that determine its evolution and fate.

p. 36 The standard cosmological model
p. 39 ESSAY: Dan Hooper
The mystery of dark matter
p. 40 Seven ways to make dark matter
p. 42 Have we got gravity wrong?
p. 43 What is dark energy?
p. 45 The Hubble tension
p. 46 Beyond the cosmological principle
p. 48 INTERVIEW: Katie Mack
Contemplating the fate of the universe
p. 49 Big crunch, big freeze or big rip?



CHAPTER 4

BLACK HOLES

Few predictions of Einstein's relativity excite the imagination like black holes, objects so massive not even light can escape them. Only since the 1960s have we been convinced they exist – and even now we aren't entirely sure what we have seen.

p. 54 The reality of black holes

p. 56 The first black hole image

p. 57 ESSAY: Paul Davies
The great black hole paradox

p. 61 What happens if you fall into a black hole?

p. 62 What made black holes so big?

p. 64 INTERVIEW: Andrea Ghez
How I proved supermassive black holes are real

p.66 ESSAY: Carlo Rovelli
Black hole... white hole?

CHAPTER 5

GRAVITATIONAL WAVES

For a long time, the existence of the minuscule ripples in space-time known as gravitational waves were the only prediction of general relativity that hadn't been directly verified. That all changed with a stunning detection by the LIGO collaboration in 2015, almost exactly 100 years after Einstein announced the theory.

p. 72 The long road to detection

p. 75 The gravitational wave background

p. 77 PICTURE ESSAY: How we found gravitational waves

p. 81 How does LIGO work?

CHAPTER 6

BEYOND RELATIVITY

Einstein's relativity has proved to be a peerless guide to the universe for more than a century. Yet it has deficiencies and limitations that have cosmologists and physicists wishing for more. Chief among their beefs is general relativity's failure to mesh with quantum theory – a roadblock on the route to a theory of everything.

p. 84 Theories of everything

p. 87 Six routes to a theory of everything

p. 88 Beyond the big bang

p. 93 ESSAY: Sean Carroll
Unpicking the fabric of the universe

CHAPTER 1

BASICS OF RELATIVITY

Albert Einstein's contributions to modern physics are unmatched.
His reputation rests above all on a theory he presented to the Prussian Academy of Sciences in Berlin in a series of lectures in the autumn of 1915.

The general theory of relativity is, at its heart, a theory of gravity. But in explaining how the force that sculpts large-scale reality works through the warping of space and time, it revolutionised our view of the cosmos.

The origins of that revolution lie earlier, however. In 1905, Einstein had taken an assumption about light's always-constant speed baked into James Clerk Maxwell's theory of the electromagnetic force to show how two observers in motion relative to each other will perceive space and time differently.

This is what became known as the special theory of relativity – and it is the natural starting point for any exploration of Einstein's universe.

THE SPECIAL THEORY OF RELATIVITY

Published in Einstein's "annus mirabilis" of 1905, special relativity works through the consequences of light always travelling at the same speed, showing how space, time, mass and energy warp to accommodate it. It was an insight that upended all our intuitions, and it still bends minds today.

PREVIOUS PAGE: JUST_SUPER/ISTOCK
BEHOLDINGEYE/ISTOCK

THE basic scientific statement of relativity goes back to way before Einstein, to the time of Galileo Galilei and Isaac Newton. It holds simply that the same laws of physics apply for all "inertial" observers – ones moving in straight lines at constant speeds relative to one another. Galileo wrote of a passenger inside a ship who cannot tell if it is moving or standing still "so long as the motion is uniform and not fluctuating this way and that".

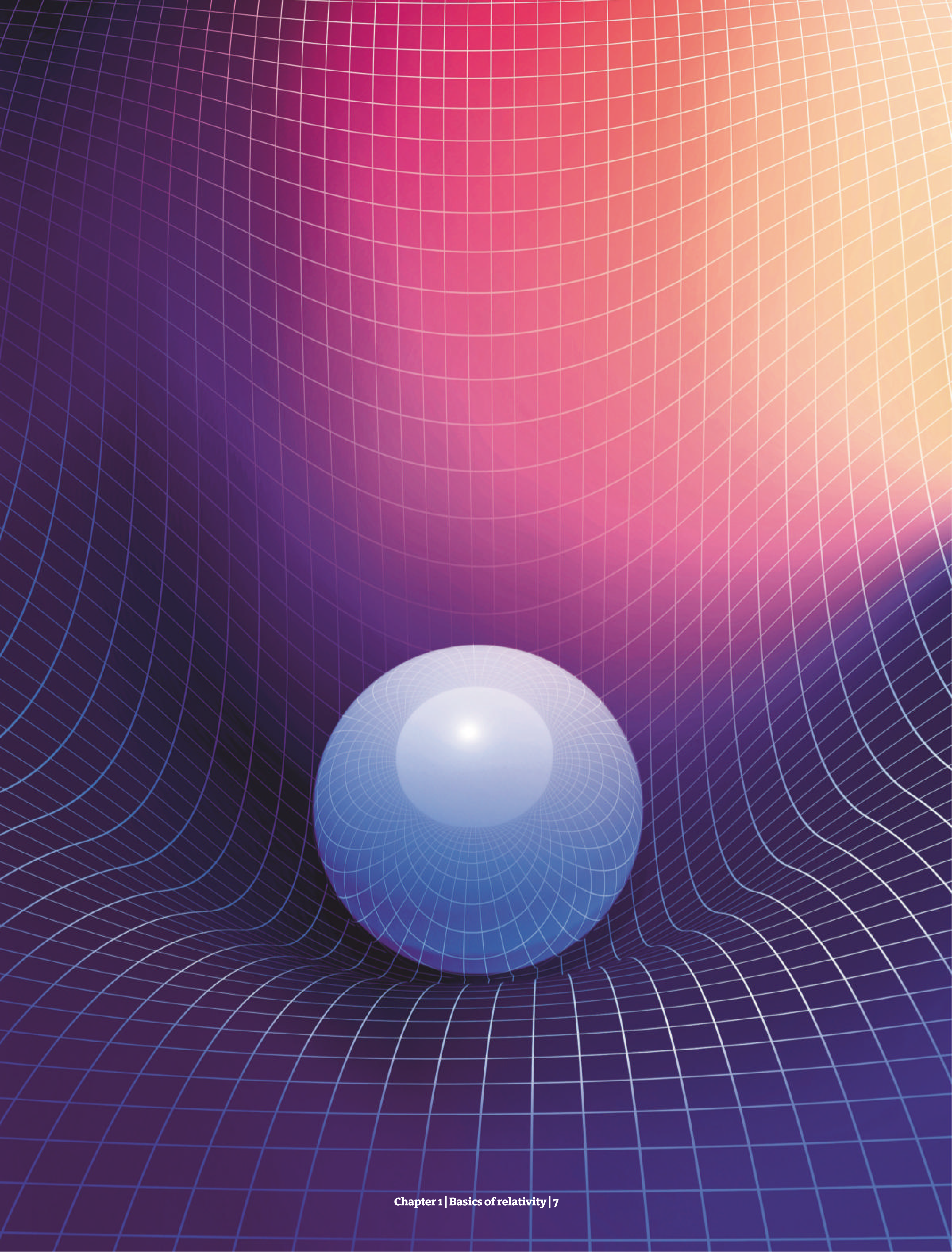
The analogy appeared in Galileo's *Dialogue Concerning the Two Chief World Systems*, his treatise of 1632 that got him into hot water with the Catholic church for discussing Copernicus's idea that Earth goes round the sun. It was aimed squarely at those sceptics who believed that Earth couldn't be moving because they couldn't feel it.

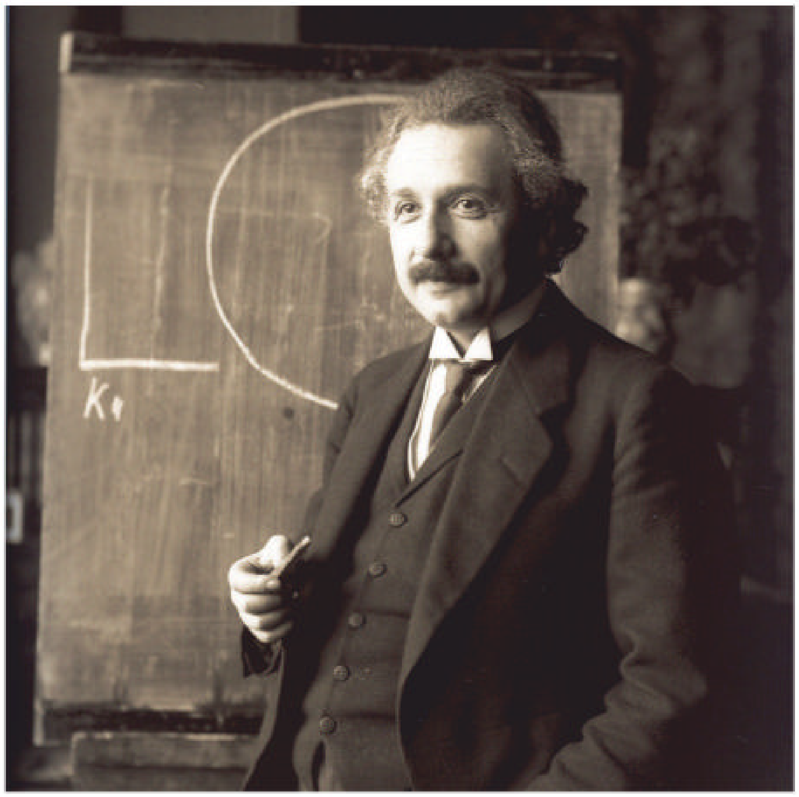
Einstein's revolutionary insight lay in applying relativity to light's strange behaviour. The speed of light, usually denoted as c, had been measured for the first time in 1676 by Danish astronomer Ole Rømer, based on observations of the timing of eclipses of the moons of Jupiter. By the end of the 19th century, it had been determined reasonably accurately to be just under 300 thousand kilometres a second, or 3×10^8 metres per second).

During the 1860s, Scottish physicist James Clerk Maxwell developed his mathematical representation of the way in which electromagnetic disturbances propagate as waves. Maxwell's equations included a constant that specified the speed with which electromagnetic waves move. That speed is precisely the speed of light. Maxwell had proved that light is a form of electromagnetic wave.

This raises a puzzle about the nature of light. Imagine two people, one on a moving train and one stationary on a platform. If the train is moving at 100 kilometres per hour and the person on the platform rolls a ball

›





Einstein's theories of relativity remain his monuments

GL ARCHIVE/ALAMY STOCK PHOTO

along it at a speed of 1 kilometre per hour in the same direction as the train, the observer will measure the speed of the ball as 99 kilometres per hour, receding into the distance. Here, measured velocity is relative to the motion of the observer.

But what is the one universal light speed in Maxwell's equations measured relative to? During the late 19th century, physicists thought that there must be an "ether", an invisible, all-pervasive fluid in which light waves move, just as sea waves move across the surface of the water. They reasoned that the unique speed of light defined by Maxwell's equations referred to its speed through the ether, and expected that observers moving through the ether in different ways would measure different speeds of light.

No experiment ever showed any such effect. In one famed test, physicist Albert Michelson and chemist Edward Morley, working in the US, showed that the speed of light measured from Earth was the same at different times of year, when it was facing different ways in its orbit of the sun, and so moving in different directions in respect to the hypothetical ether.

In what became known as the special theory of relativity (but only after the general theory was formulated), Einstein took this idea of a constant light speed and ran with it. The strangeness of the resulting picture can be seen by imagining another simple experiment involving a train.

If a person on a train sets up a source of light in the middle of a carriage, then sends off two pulses of light in opposite directions at the same time, they will see them reach the end walls of the carriage simultaneously. But because the train is moving forwards, if light speed is always the same, an outside observer sees the rearward-moving pulse hit the back wall of the carriage before the forward-moving pulse hits the front wall. In other words, events that happen at the same time for one observer can occur at different times for an observer who is moving relative to the first observer. Simultaneity is relative.

If observers who move relative to one another can't agree on the simultaneity of events, they can't agree on the outcome of measurements involving time. You can use simple calculations of the geometry of scenarios similar to those sketched above to show that, if a clock is moving at a speed v, time is dilated – lengthened – by exactly a factor $1/(1-v^2/c^2)^{\frac{1}{2}}$, an expression which appears in many relativistic calculations. Moving clocks tick slower. Do the same calculations with length and you find that moving objects will be seen as shorter: a moving stick is shrunk in the direction of its motion by a factor $(1-v^2/c^2)^{\frac{1}{2}}$.

The nature of this factor means that it only becomes significantly different from one when the speed of the moving object v becomes close to the speed of light, c, which is why we don't notice relativistic effects in everyday life. But they are universal. In 1971, two US physicists, Joseph Hafele and Richard Keating, confirmed time dilation by transporting very precise caesium-beam clocks around the world on commercial jets. After 45 hours, the researchers found their clocks ran slower than those on Earth. You age a fraction of a second less while in flight than you would if you stayed at home.

The relativity of stuff doesn't stop there, either. As Einstein concluded in a paper published in 1906, mass is also subject to its rules. The mass of a moving object is equal to its mass when stationary divided by the familiar relativistic factor, $(1\text{-}v^2/c^2)^{\frac{1}{2}}$. Moving objects have more mass than objects at rest.

One way of interpreting this increase in mass as velocity increases is in terms of kinetic energy. The mass of the moving object exceeds its rest mass by an amount equal to its kinetic energy divided by c^2. This led Einstein to infer that the rest mass itself is equivalent to an energy E/c^2. In other words, $E = mc^2$, his famous equation of mass-energy equivalence.

Strange though the predictions of the special theory of relativity are, they have all been borne out by experiment – not least in the conversion of rest mass into energy in the explosion of an atomic bomb. It was only by incorporating special relativity into the new quantum description of particles, also a product of the early decades of the 20th century, that British physicist Paul Dirac developed a satisfactory mathematical description of the behaviour of the electron and other subatomic particles in 1928. This relativistic version of quantum mechanics provides our understanding of the behaviour of electrons in atoms, and the way in which they occupy stable "shells" around the nucleus – the basis for modern chemistry.

A little later, by merging special relativity and quantum mechanics with electromagnetism, physicists developed a theory known as quantum electrodynamics, or QED. This supersedes Maxwell's classical theory of electromagnetism to give a complete quantum description of the way charged particles interact. That success led to the construction of a similar theory to describe the behaviour of quarks (the particles that make up the protons and the neutrons of the atomic nucleus), known as quantum chromodynamics, or QCD. Together, these two theories make up the "standard model" of particle physics, the theory that underpins our understanding of all the forces of nature today, bar gravity.

Were special relativity Einstein's sole contribution to physics, he would be numbered among the greats. But the theory was a mere hors d'oeuvre. As we said earlier, special relativity only applies to inertial observers that are at rest or moving at constant speeds. It was by generalising its lessons to far more involved questions of how time, space and other properties of the physical world look relative to one another when observers are accelerating that Einstein developed the theory for which he is revered – and that underpins our understanding of the universe today.

→ Chapter 6 has more on quantum theories of gravity

ALBERT EINSTEIN: A LIFE

Albert Einstein's name has become a byword for genius. He is arguably the most influential and recognisable scientist of the 20th century.

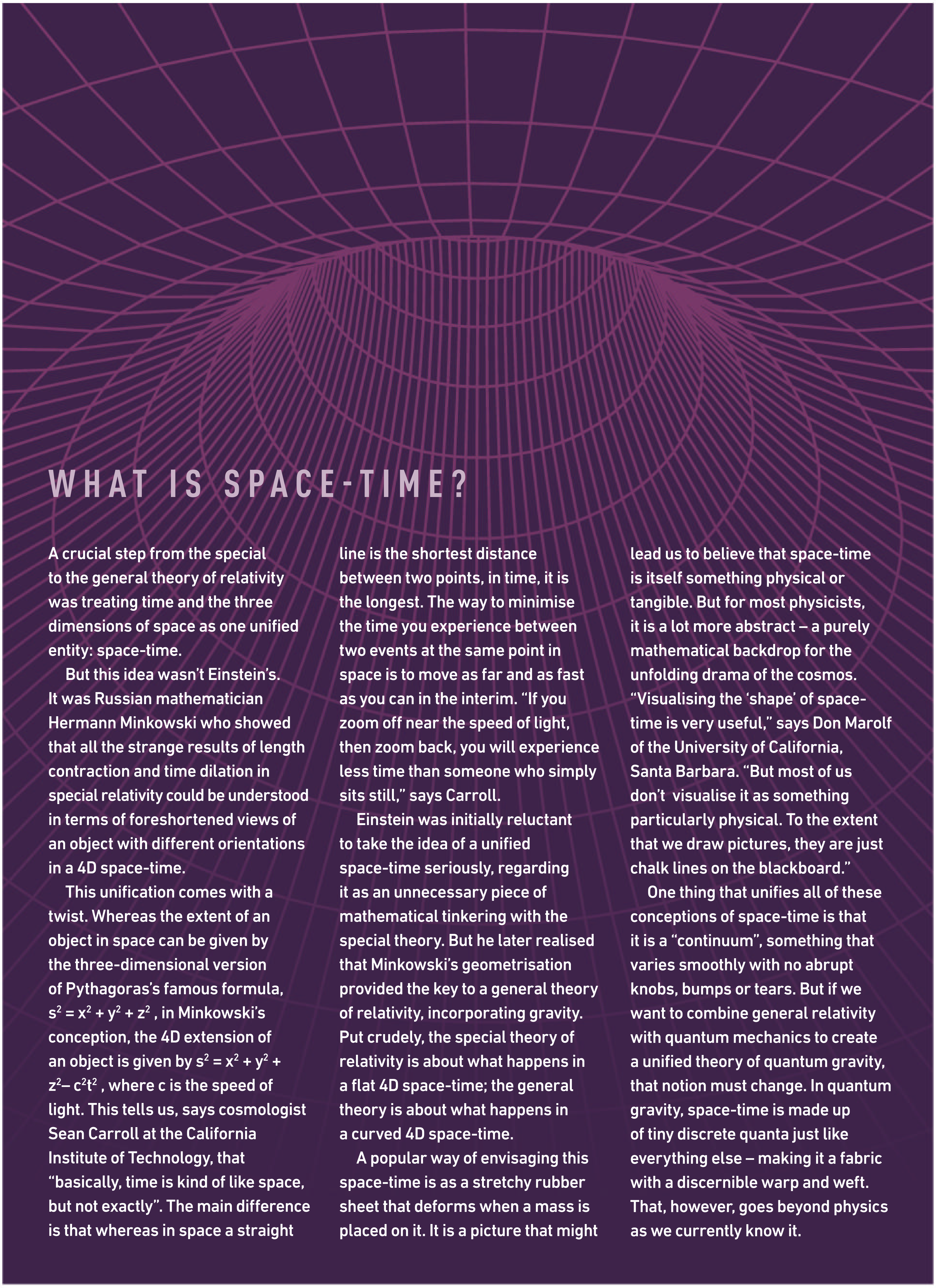
Einstein was born into a family of Jewish heritage in the city of Ulm in what is now the south of Germany on 14 March 1879. He moved to Switzerland in 1895, becoming a Swiss citizen in 1901. Einstein studied physics and maths in Zurich, receiving a PhD from the university there in 1905, by which stage he was working at the Swiss Patent Office in Bern.

It was in this year, at the age of 26, that Einstein published three papers which would alone have ranked him among the half-dozen great pioneers of 20th-century physics. Those papers dealt with the special theory of relativity, the way tiny particles move through the air or a liquid (Brownian motion) and how light is quantised into packets of energy, photons, when emitted from certain metals (the photoelectric effect).

This last contribution was singled out in Einstein's citation for the Nobel prize in physics in 1921 – although he almost didn't get it at all, the prize-awarding committee having tided over his nomination from the previous year because they had been unsure his achievements merited the award. That might have had something to do with the controversy generated by what is now viewed as his greatest single contribution, the general theory of relativity, which he presented in 1915.

Einstein remained active in debates about the fundamental nature of reality for the rest of his life. Despite having been one of the pioneers of quantum theory, he became a notable sceptic of the theory as it developed during the 1920s and 1930s – yet he spent much of his last decades engaged in a fruitless effort to unify it with general relativity.

On the accession of the Nazis to power in Germany in 1933, Einstein, by then working at the Kaiser Wilhelm Institute for Physics in Berlin, decided to settle in the US, spending the rest of his working life at the Institute for Advanced Study in Princeton, New Jersey. He died in New Jersey on 18 April 1955, at the age of 76.



IHOR LISHCHYSHYN/ISTOCK

WHAT IS SPACE-TIME?

A crucial step from the special to the general theory of relativity was treating time and the three dimensions of space as one unified entity: space-time.

But this idea wasn't Einstein's. It was Russian mathematician Hermann Minkowski who showed that all the strange results of length contraction and time dilation in special relativity could be understood in terms of foreshortened views of an object with different orientations in a 4D space-time.

This unification comes with a twist. Whereas the extent of an object in space can be given by the three-dimensional version of Pythagoras's famous formula, $s^2 = x^2 + y^2 + z^2$, in Minkowski's conception, the 4D extension of an object is given by $s^2 = x^2 + y^2 + z^2 - c^2t^2$, where c is the speed of light. This tells us, says cosmologist Sean Carroll at the California Institute of Technology, that "basically, time is kind of like space, but not exactly". The main difference is that whereas in space a straight line is the shortest distance between two points, in time, it is the longest. The way to minimise the time you experience between two events at the same point in space is to move as far and as fast as you can in the interim. "If you zoom off near the speed of light, then zoom back, you will experience less time than someone who simply sits still," says Carroll.

Einstein was initially reluctant to take the idea of a unified space-time seriously, regarding it as an unnecessary piece of mathematical tinkering with the special theory. But he later realised that Minkowski's geometrisation provided the key to a general theory of relativity, incorporating gravity. Put crudely, the special theory of relativity is about what happens in a flat 4D space-time; the general theory is about what happens in a curved 4D space-time.

A popular way of envisaging this space-time is as a stretchy rubber sheet that deforms when a mass is placed on it. It is a picture that might lead us to believe that space-time is itself something physical or tangible. But for most physicists, it is a lot more abstract – a purely mathematical backdrop for the unfolding drama of the cosmos. "Visualising the 'shape' of space-time is very useful," says Don Marolf of the University of California, Santa Barbara. "But most of us don't visualise it as something particularly physical. To the extent that we draw pictures, they are just chalk lines on the blackboard."

One thing that unifies all of these conceptions of space-time is that it is a "continuum", something that varies smoothly with no abrupt knobs, bumps or tears. But if we want to combine general relativity with quantum mechanics to create a unified theory of quantum gravity, that notion must change. In quantum gravity, space-time is made up of tiny discrete quanta just like everything else – making it a fabric with a discernible warp and weft. That, however, goes beyond physics as we currently know it.

YULKAPOPKOVA/ISTOCK

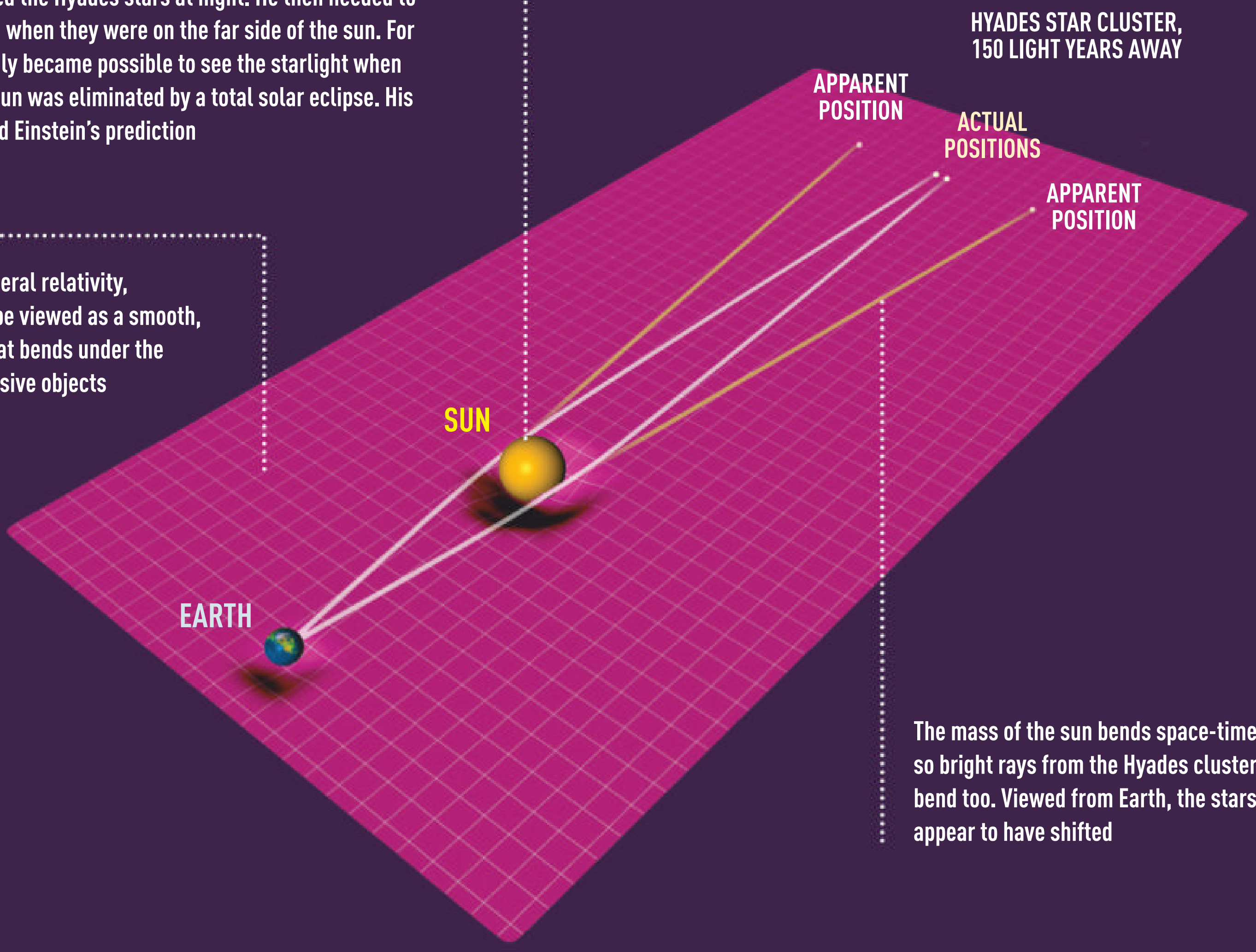
General relativity can be summed up in just 12 words: "Space-time tells matter how to move; matter tells space-time how to curve". But this description from physicist John Wheeler hides a more complex and profound truth about the nature of gravity – one Einstein only grasped by considering what happens when it isn't there.

THE GENERAL THEORY OF RELATIVITY

EINSTEIN used to say that the unique flash of insight that set him on the path to the general theory of relativity came when he realised that a man falling from a roof – or a person trapped inside a freely falling lift – doesn't feel the force of gravity. People in the falling lift will float, completely weightless, able to push themselves from wall to wall or floor to ceiling with great ease.

It is a picture that is now familiar from pictures of astronauts in spacecraft, falling freely in orbit around Earth. In weightless conditions, objects precisely obey the laws of motion that Isaac Newton devised in the 17th century, proceeding in straight lines unless interfered with by forces. Einstein had to imagine all the things we have seen for ourselves in footage from space: pencils hanging in mid-air, liquids that refuse to pour and so on. His genius saw all this, and what everybody else missed. If the acceleration of a falling lift, plunging downwards at an ever-increasing speed, can precisely cancel out the force of gravity, then that force and acceleration are exactly equivalent to one another.

›



Albert Einstein suggested that light rays skimming past the sun would be bent by its gravity. To test the idea, Arthur Eddington first photographed the Hyades stars at night. He then needed to photograph them when they were on the far side of the sun. For this picture, it only became possible to see the starlight when the glare of the sun was eliminated by a total solar eclipse. His images confirmed Einstein's prediction

According to general relativity, space-time can be viewed as a smooth, flexible sheet that bends under the influence of massive objects

The mass of the sun bends space-time, so bright rays from the Hyades cluster bend too. Viewed from Earth, the stars appear to have shifted

EDDINGTON'S EXPERIMENT

In 1919, four years after Einstein first presented general relativity to the world, British astronomer Arthur Eddington travelled to the island of Príncipe off the coast of West Africa to see if he could detect the lensing of light predicted by the theory.

Eddington's plan was to observe a bright cluster of stars called the Hyades as the sun passed in front of them, as seen from Earth. To see the starlight, he needed a total solar eclipse to blot out the sun's glare. If Einstein's theory was correct, the positions of the stars in the Hyades would appear to shift by about 1/2000th of a degree.

To pinpoint the position of the Hyades in the sky, Eddington first took a picture at night from Oxford, UK. On 29 May 1919, he photographed the Hyades as they lay almost directly behind the sun during the total eclipse seen from Príncipe that day. Comparing the measurements, Eddington showed the shift was as Einstein had predicted and too large to be explained by Newton's theory.

Following the eclipse expedition, there was some controversy that Eddington's analysis had been biased towards general relativity. Matters were put to rest in the late 1970s when the photographic plates were analysed again and Eddington's analysis was shown to be correct.

Eddington's result turned Einstein into an international superstar. "LIGHTS ALL ASKEW IN THE HEAVENS; Men of Science More or Less Agog Over Results of Eclipse Observations. EINSTEIN THEORY TRIUMPHS" was the headline of the story in *The New York Times*.

General relativity has never failed an experimental test in the century since. As more and more consequences of his theory have been discovered, from the expanding universe to the existence of black holes and gravitational waves, its hold on public imagination hasn't diminished either.

“Einstein constructed a picture in which gravity is just a form of acceleration”

This is actually no different from a principle noted more than three centuries earlier by Galileo: that falling objects accelerated at the same rate regardless of their mass. Famously, a feather and a hammer dropped from the Leaning Tower of Pisa will hit the ground at the same time, once you discount air resistance. During the Apollo 15 lunar landing in 1971, astronaut David Scott confirmed this principle on the airless moon.

Newton showed that this could only be true if an odd coincidence held: inertial mass, which quantifies a body’s resistance to acceleration, must always equal gravitational mass, which quantifies a body’s response to gravity. There is no obvious reason why this should be so, yet no experiment has ever prised these two quantities apart. This was the nub of what became the general theory. In the same way that he had used light’s constant speed to construct the special theory of relativity, Einstein declared the sameness of inertial and gravitational mass to be a principle of nature: the equivalence principle.

The power of this insight is clear if we imagine the lift replaced by a closed laboratory that is being accelerated through space by a constant force. Everything in the laboratory falls to the floor, and a physicist who carries out experiments inside will be unable to tell whether the downward force is due to an acceleration or to the force of gravity pulling things down. We are used to thinking of acceleration as being caused by a force, but from the point of view of the lift’s occupant, that force is caused by acceleration (see diagram, right).

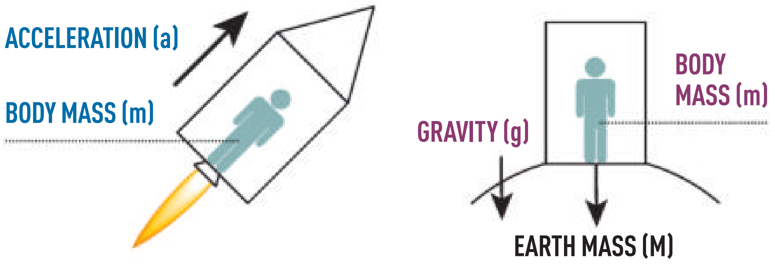
Armed with the equivalence principle, plus the new conception of space and time as an interwoven space-time, Einstein constructed a picture in which gravity is just a form of acceleration. Massive objects bend space-time around them, making things

›

Einstein’s equivalence principle states that the physics of acceleration and gravity work in exactly the same way. But there is no reason why that should be the case

Accelerate a rocket in gravity-free space and a body’s inertial mass will resist the motion

The mutual attraction between gravitational masses is what keeps our feet on the ground



In all situations

inertial mass ≡ gravitational mass

acceleration ≡ gravity

It’s only because the equivalence principle appears to be true that bodies at the same distance from Earth fall to the ground at the same rate

NEWTON’S 2ND LAW OF MOTION

NEWTON’S GRAVITATIONAL LAW

$$\text{force} = m \times \underset{\text{acceleration}}{a} = \underset{\text{mass}}{m} \times \frac{\overset{\text{Gravitational constant}}{G}\ \overset{\text{Mass of Earth}}{M}}{\underset{\text{Distance between body and centre of Earth}}{r^2}}$$

If $m = m$

then $\underbrace{a = \frac{GM}{r^2}} = g$

At a distance (r) from Earth’s centre, the acceleration due to gravity (g) is ALWAYS THE SAME

NEWTON VS EINSTEIN

In 1686, Isaac Newton proposed an incredibly powerful theory of motion. At its core was the law of universal gravitation, which states that the force of gravity between two objects is proportional to each of their masses and inversely proportional to the square of their distance apart. Newton's law is universal because it can be applied to any situation where gravity is important: apples falling from trees, planets orbiting the sun and many, many more.

For more than 200 years, Newton's theory of gravity was successfully used to predict the motions of celestial bodies and accurately describe the orbits of the planets in the solar system. Such was its power that in 1846, French astronomer Urbain Le Verrier was able to use it to predict the existence of Neptune.

There was, however, one case where Newton's theory didn't seem to give the correct answer. Le Verrier measured Mercury's orbit with exquisite precision and found that it drifted by a tiny amount – less than one-hundredth of a degree over a century – relative to what would be expected from Newton's theory. As Mercury follows an elliptical orbit around the sun, the ellipse itself shifts slightly each orbit (or "precesses"), tracing out a pattern like a child's drawing of the petals of a daisy. Le Verrier thought that this indicated the presence of another unseen planet, which he called Vulcan, hidden to us by the sun.

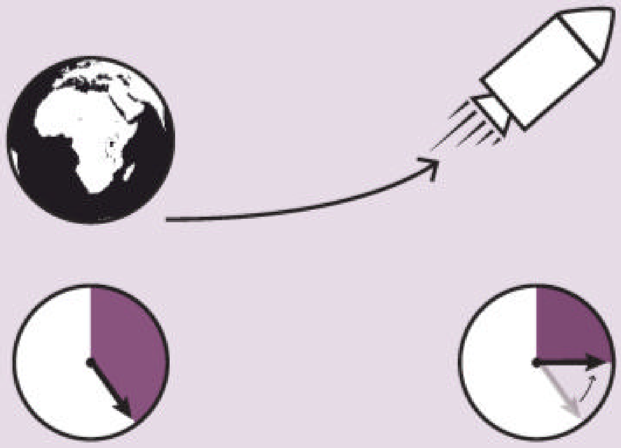
The discrepancy between Newton's theory and Mercury's orbit was still unresolved at the beginning of the 20th century. General relativity resolved it – and proved Newton's theory was just an approximation to something bigger. Anywhere that gravity is weak, general relativity and Newton's inverse square law give the same answers to a certain level of precision. But in a strong field, such as close to the sun, or where greater precision is needed – for example, in pinpointing locations using today's GPS satellites – general relativity must be used.

Einstein's theories of relativity – special and general – encompass two effects that influence our perception of space and time

SPECIAL RELATIVITY

Moving clocks run slower

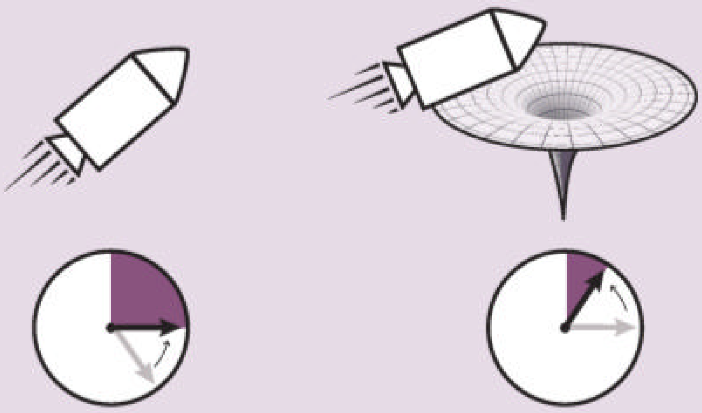
To an outside observer, the crew of a spaceship travelling close to the speed of light will age less than people on Earth



GENERAL RELATIVITY

Clocks run slower in high gravity

A spaceship entering a high gravitational field like that of a black hole will experience even less time



F. W. DYSON/A. S. EDDINGTON/C. DAVIDSON (1920)

Arthur Eddington's 1919 solar eclipse expedition confirmed relativity

appear to accelerate towards them. That explains why we feel a downwards pull towards Earth and why Earth orbits the sun. Objects are simply following a path of least resistance, a geodesic – the equivalent of a straight line through a curved portion of space-time.

General relativity applies to everything in the cosmos, including light. It means that all the time- and space-warping effects of movement embodied in the special theory are also in the gift of gravity, too. Most notably, clocks run slower in higher gravity – a clock at sea level on Earth ticks slower than one at the summit of Mount Everest. And sometimes the effects of special and general relativity play against one another. Astronauts on the International Space Station (ISS), for example, age a little less because of the velocity at which they travel (special relativity) and a little more because they enjoy less of the gravity of Earth (general relativity). Velocity wins out, leaving each ISS astronaut who completes a six-month tour of duty 0.007 seconds younger than someone who stayed on Earth.

The theory of general relativity as presented by Einstein in a series of lectures to the Prussian Academy of Sciences in Berlin in October and November 1915 came in a series of fearsomely complicated "field equations". They used mathematical techniques of tensor algebra that were entirely new at the time to relate the distribution of mass and energy in the universe to the local curvature of space-time. Working through the implications of the mathematics would take time, provoking bitter controversy over decades in certain cases.

But some concrete predictions were already clear: for example, the idea that a beam of light should be bent by gravity when it passes near the sun in exactly the same way as light is bent when seen by a physicist in an accelerated frame of reference. One consequence was that stars that were observed "near" the sun in the sky during a total eclipse (but were actually, of course, much further away along the line of sight) would be displaced by a certain amount compared with their observed positions at other times. The prediction was confirmed by observations made by British physicist Arthur Eddington on the West African island of Príncipe during a total solar eclipse in 1919 (see "Eddington's experiment", page 12).

Unlike all of Einstein's previous work, including on special relativity, the general theory of relativity wasn't a response to any observational puzzle. Einstein was motivated by a deeper philosophical need, the quest for simplicity and unity in nature. If it hadn't been for Einstein, a comprehensive theory of gravity might not have been developed for decades, until, in the 1960s and 1970s, scientists were pressed to consider the need for such a theory by the discovery of objects such as black holes, pulsars and quasars. As it is, all general relativity's many, often surprising consequences, from the expanding big bang universe to the existence of dark matter, dark energy and gravitational waves, have become part of the human intellectual canon over the past century. It is, quite simply, a towering intellectual achievement. ▮

→
Chapter 2 has more on the big-bang universe

→
Chapter 3 has more on dark matter and energy

→
Chapter 4 has more on black holes

→
Chapter 5 has more on gravitational waves

EINSTEIN'S UNIVERSE: A TIMELINE

From its origins as the work of a lone genius in the early years of the 20th century, relativity has developed into the basis of an entire model of the universe. Phenomena that it predicts, such as black holes and gravitational waves, have been shown to be real – while other mysteries it has called into life, such as the nature of dark matter and dark energy, remain unsolved.

1905

In a paper titled "On the Electrodynamics of Moving Bodies", Albert Einstein proposes new rules to calculate the relative motion of objects travelling close to the speed of light, the cosmic speed limit – what later becomes known as the special theory of relativity

1915

After 10 years of work, Einstein generalises the rules of special relativity to create a new theory of gravity as a product of mass warping a new unified space-time. He presents his field equations of general relativity to the Prussian Academy of Sciences in Berlin

1916

Einstein uses general relativity to predict the existence of gravitational waves, ripples in space-time produced when massive bodies interact

MIKHAIL KONOPLEV/ISTOCK

1917

Einstein introduces an extra term into his equations, the cosmological constant, to balance out gravity and produce a static universe that is neither expanding nor contracting

1919

Arthur Eddington observes the sun's mass bending light during an eclipse over the island of Príncipe – an effect known as gravitational lensing that was predicted by Einstein

1920s

Alexander Friedmann and Georges Lemaître independently find a solution to Einstein's equations that describes a uniformly expanding universe

→ Page 20

1920s

Edwin Hubble and others show that far-off galaxies are moving away from us – the first hint of an expanding "big bang" universe. Einstein decries his cosmological constant as his "greatest blunder"

1930

Subrahmanyan Chandrasekhar shows that certain massive stars could collapse into bodies so dense that no light could escape from them. These later become known as black holes

→ Page 54

1933

Fritz Zwicky observes that galaxies in clusters are seemingly being whirled around by the gravity of invisible matter – the first hint of the existence of dark matter

→ Page 36

1940s

Theorists predict that if the universe is expanding from a hot and dense beginning in a big bang, it should have left behind an afterglow: the cosmic microwave background

→ Page 23

1964

The cosmic microwave background is accidentally discovered by Arno Penzias and Robert Wilson as unexplained noise in a radio antenna, kicking off relativity's "golden age"

1970s

Vera Rubin provides convincing evidence that most galaxies contain dark matter, which is causing them to rotate faster

1972

X-ray emissions from a body known as Cygnus X-1 provide the first evidence of a star's collapse into a stellar-mass black hole

1974

Russell Hulse and Joseph Taylor discover a pair of neutron stars whose orbits are slowing exactly as if they are losing energy by emitting gravitational waves

1974

Stephen Hawking shows theoretically that quantum effects can cause black holes to evaporate, emitting Hawking radiation – posing the question of what happens to the matter they swallow → Page 57

1980

Alan Guth and others propose that the big-bang universe was smoothed out by the universe undergoing a period of breakneck expansion – inflation – in its first instants → Page 30

1989

NASA launches COBE, a satellite to study the cosmic microwave background. It reveals a largely homogeneous radiation field, supporting the idea of an inflationary big bang

1998

Studies of far-off supernovae surprisingly reveal that the universe's expansion is accelerating. Einstein's cosmological constant is revived as one identity for the "dark energy" causing this effect → Page 43

2000s

Ever-more detailed studies of the cosmic microwave background support the picture of a cosmos that began in an inflationary big bang dominated by dark matter and dark energy

2003

The DAMA experiment under the Gran Sasso mountain in central Italy claims to have seen a signal of Earth ploughing through a sea of dark matter – but other experiments fail to verify it

2008

The Large Hadron Collider fires up at CERN near Geneva, Switzerland. One of its aims is to make dark matter particles

2014

Physicists working with the BICEP2 telescope at the South Pole say they have seen the imprint of primordial gravitational waves on the cosmic microwave background – a claim later retracted

2016

A century after Einstein predicted them, the Laser Interferometer Gravitational-Wave Observatory (LIGO) announces the first direct detection of a gravitational wave passing through Earth. The wave, caused by a merger of two black holes, was sensed on 14 September 2015 → Page 72

2019

The Event Horizon Telescope collaboration publishes the first image of a black hole, a supermassive giant at the heart of galaxy Messier 87, 53 million light years away

CHAPTER 2

THE BIG BANG UNIVERSE

The most striking implications of the general theory of relativity came in its predictions for the universe as a whole. Shortly after Albert Einstein published his theory, Alexander Friedmann and Georges Lemaître independently showed that the entire cosmos should evolve in response to all the energy it contains, starting off small and dense and expanding and diluting with time.

Einstein was initially sceptical of Friedmann and Lemaître's conclusions, favouring an eternal, unchanging, static universe. A discovery by astronomer Edwin Hubble in the late 1920s proved to be the turning point. What he found led us to today's big bang theory – the idea that our observable universe, at least, began in a blip of unimaginable heat and density some 13.8 billion years ago.

THE EXPANDING UNIVERSE

The idea that the universe was anything other than a static, timeless entity with no beginning and no end had occurred to few physicists by the early 20th century. Only a unique confluence of a new theory and compelling new observations forced them to revisit their assumptions.

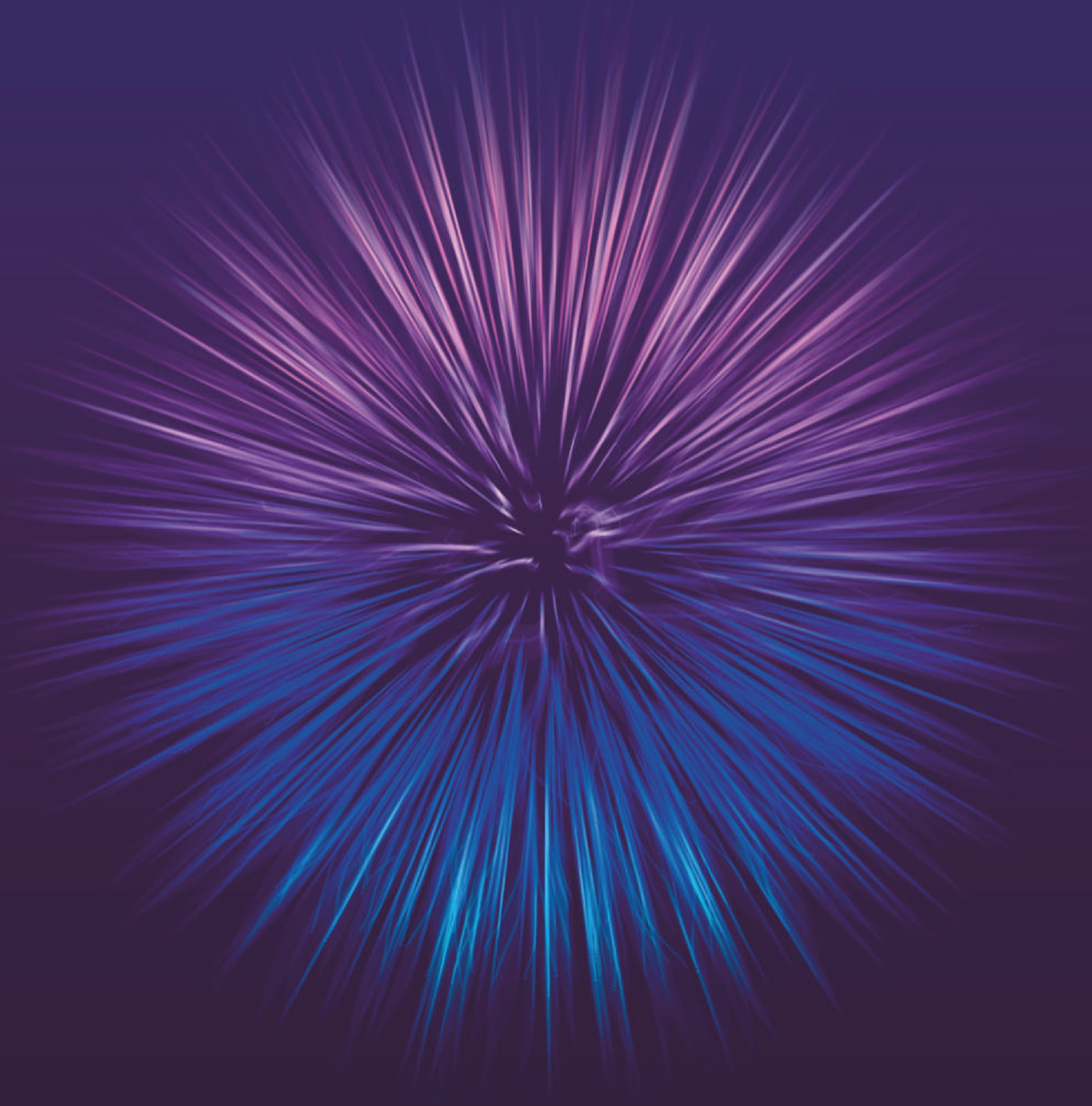
PREVIOUS PAGE: JUST_SUPER/ISTOCK
STUDIOM1/ISTOCK

WHY are we here? How did the universe begin? According to the Bushongo people of central Africa, before us there was only darkness, water and the god Bumba. One day Bumba, in pain from a stomach ache, vomited up the sun. The sun evaporated some of the water, leaving land. Still in discomfort, Bumba vomited up the moon, the stars and then the leopard, the crocodile, the turtle and, finally, humans.

This creation myth, like many others, wrestles with the kinds of questions that we all still ask today. With the general theory of relativity, Albert Einstein provided the tools to supply, for the first time, some sort of scientifically grounded answer.

Not that Einstein believed it when he first saw it. At the time he was formulating the general theory, the received wisdom was that our Milky Way galaxy was the entire universe, a stable collection of stars. But when Einstein tried to describe the simplest possible mathematical model of the universe using

›



his new equations, they refused to produce such a picture. They insisted that the universe must either be expanding or contracting.

The only way Einstein could hold the model universe still, to mimic the appearance of the Milky Way, was to add an extra term to the equations, which he called the cosmological constant. In 1917, he wrote "that term is necessary only for the purpose of making possible a quasi-static distribution of matter, as required by the fact of the small velocities of the stars". It was an artificial addition that he would later decry as his "greatest blunder".

→
See page 43 for how the cosmological constant might be reborn as dark energy

In the 1920s, Russian mathematician Alexander Friedmann and Belgian priest-cosmologist Georges Lemaître independently reverted to the simple solution contained in Einstein's own equations: that the universe was expanding. Extrapolating backwards in time, Lemaître reasoned that it ought to have started off as a small "primeval atom".

Einstein wasn't impressed, arguing that though the solution was mathematically sound, it wasn't feasible physically. "Your calculations are correct, but your grasp of physics is abominable", he is supposed to have said to Lemaître.

In fact, the theory had just been ahead of the observations. In 1923, US astronomer Edwin Hubble began to make observations with a telescope on Mount Wilson in California, and proved that the Milky Way was just one galaxy among thousands of millions of others scattered throughout space. As he examined these galaxies further, to his surprise, he found that nearly all of them were moving away from us. Moreover, the more distant the galaxies, the faster they were moving away.

That had potentially huge repercussions. If galaxies are moving apart in the present, they must therefore have been closer together in the past. If their speed had been constant, then they would all have been on top of one another billions of years ago. Did this imply that the universe had a beginning?

At that time, many scientists were unhappy with this notion because it seemed to imply that physics had broken down. You would have to invoke an outside agency, which for convenience you could call god, to determine how the universe began. They began to advance theories in which the universe was expanding in the present, but didn't have a beginning.

Perhaps the best known came in 1948, when Hermann Bondi, Thomas Gold and Fred Hoyle proposed the steady-state theory. Space is expanding at a constant rate but, at the same time, matter is created continuously throughout the universe. This matter is just enough to compensate for the expansion and keep the density of the universe constant. Where this matter would come from, nobody could say. But neither could the proponents of the big bang – a term, with delicious irony, originally coined by Hoyle to disparage the idea.

The steady-state theory was the principal challenger to the big bang theory for two decades. Two discoveries dealt it a fatal blow. The first came from Martin Ryle at the University of Cambridge and his colleagues. In the early 1960s, they were studying radio galaxies – enormously powerful sources of radio waves, a type of light invisible to the naked eye – and found that there were many more radio galaxies at large distances, and therefore observed further back in time, than nearby.

The excess of radio galaxies at great distances had to mean that conditions in the remote past were different from those today. But the decisive turning point in favour of the big bang theory came in 1964 – with a discovery that must count as among the most serendipitous in science. ▮

The story of universe's genesis in a hot, dense fireball was ultimately confirmed by accident. It was still a struggle to convince the doubters, cosmologist Jim Peebles recalls

HOW THE BIG BANG WAS BORN

ON 20 May 1964, Arno Penzias and Robert Wilson at the Bell Telephone Laboratories in Holmdel, New Jersey, recorded their first astronomical measurements of microwave radiation from the supernova remnant Cassiopeia A. They were using a horn antenna system first assembled in 1959 to study microwave communication – an early step in the development of today's cellphone technology.

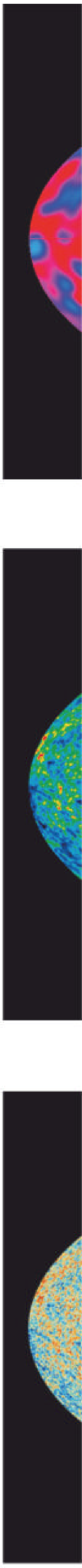
The antenna had been carefully engineered to reject radiation from the ground. But once all known sources in the sky had been painstakingly accounted for, Penzias and Wilson were left with the same problem that had been bothering Bell's engineers. The microwave sky seemed to be about 2 degrees warmer than anyone expected.

At the time, I was a young theorist just down the road in the physics department at Princeton University, in the research group of Bob Dicke. Bob was a fan of the idea that the universe began in a hot, dense state – a big bang. One idea he was exploring was that a big bang should have left behind a sea of radiation uniformly spread across the sky. Bob had set two members of his group, Peter Roll and David Wilkinson, onto building a receiver capable of detecting this radiation, and suggested I look into the theoretical

›

PROFILE
JIM PEEBLES

Jim Peebles is Albert Einstein professor of science, emeritus, at Princeton University. He was awarded one half of the 2019 Nobel prize in physics "for contributions to our understanding of the evolution of the universe and Earth's place in the cosmos".



Three satellites' views of the cosmic microwave background: COBE (active 1989 to 1993), WMAP (2001 to 2010) and Planck (2009 to 2013)

consequences of detecting or not detecting it.

The following February, I presented the idea of our search in a colloquium. A few weeks later, Bob received a phone call from Penzias. A subsequent visit to Holmdel by Bob, Peter and David convinced us that we had been scooped. The universe's origin had been discovered as an unintended consequence of a communications engineering project.

Looking back from the distance of a half-century, it is easy to forget that, back then, many physicists dismissed any investigations into how it all began as empty speculation. The sea of noise that was so troubling to Penzias and Wilson is the cosmic microwave background (CMB), which is now known to be clinching evidence for a big bang.

But it wasn't immediately seen that way, as a glance back through the pages of *New Scientist* shows. In 1976, Martin Rees, then recently elected to the illustrious Plumian professorship of astronomy at the University of Cambridge, wrote that "no plausible alternative theory" could account for the CMB's observed characteristics. Yet five years later, Rees's Plumian predecessor Fred Hoyle still saw a chance for the rival steady state theory, writing in *New Scientist* that the latest CMB measurements "differ by so much from what theory would suggest as to kill the big-bang cosmologies".

The big bang and steady state theories both have their origin in the discovery in the 1920s that distant galaxies are moving away from us, as if the universe were expanding. The big bang theory, developed in the ensuing decades, postulates that everything here now was also there back then, so the universe must be expanding from a denser early state.

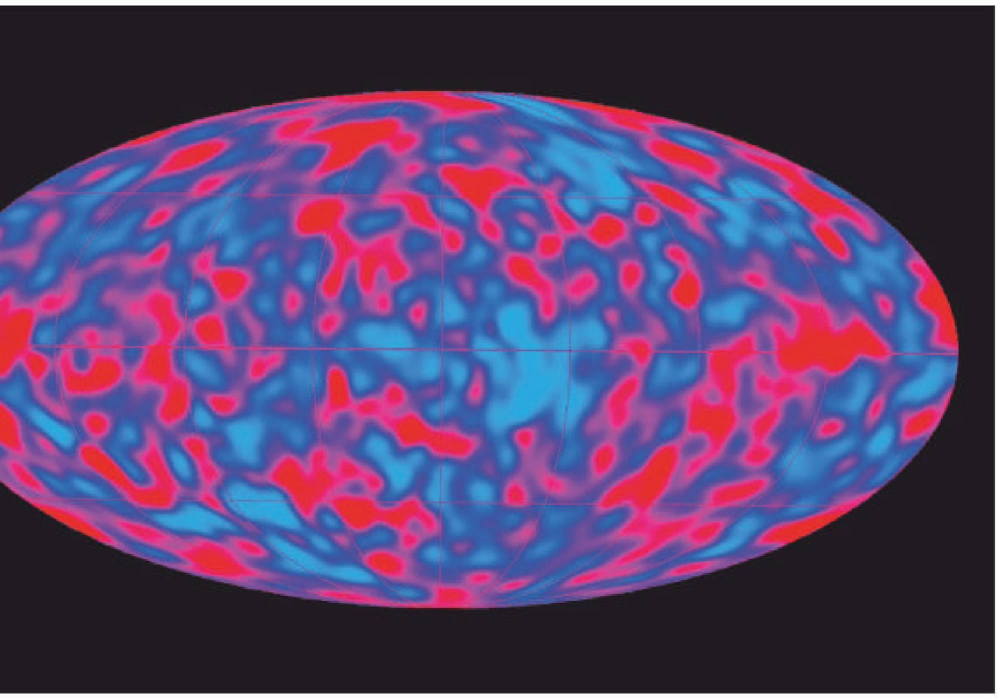
The steady state theory, which Hoyle co-devised in 1948, suggests instead that matter is continually created in the expanding universe, with new galaxies forming to fill the spaces that open up as already existing ones move apart. In this picture, the universe's past was no hotter or denser than its present.

Something like a CMB crops up in some of the earliest considerations of what a hot, dense early universe would have looked like. In the late 1930s, attempts to explain the relative abundances of the chemical elements – why is there so much more iron than gold, for instance – included the proposal that the heat of the early universe produced violent collisions that thoroughly rearranged the neutrons and protons within atomic nuclei. These analyses implicitly assumed a sea of surrounding radiation hot enough to promote the right sort of reactions. This sea would have slowly cooled as the universe expanded.

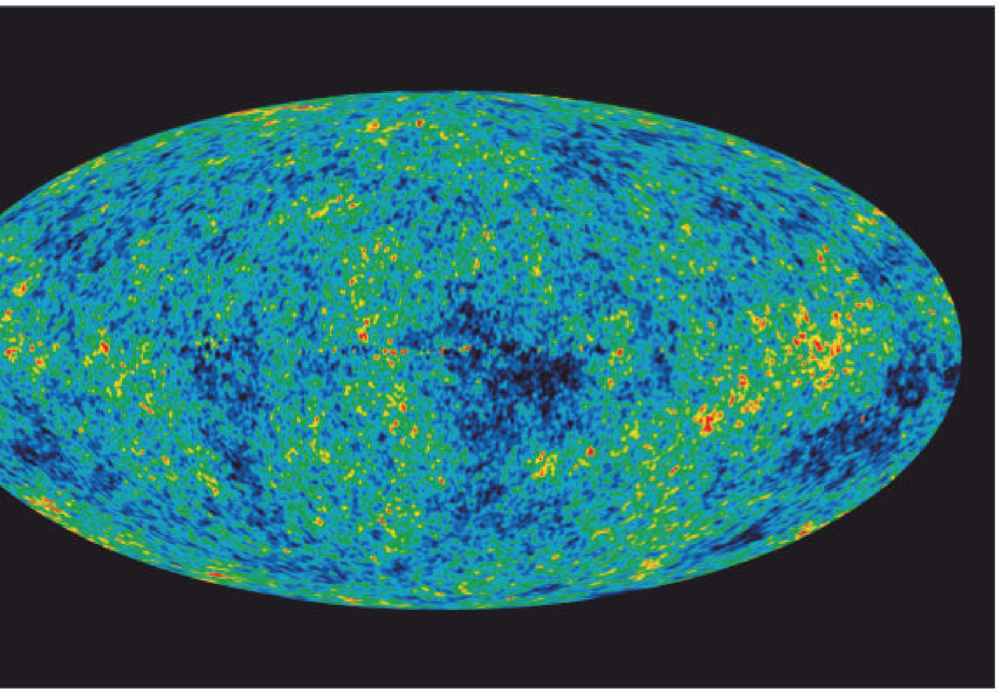
George Gamow and Ralph Alpher didn't mention this thermal radiation when they presented their famous paper on big-bang elemental abundances in 1948 – a paper celebrated both for its content and Gamow's unsolicited introduction of physicist Hans Bethe on to the author list, so it would approximate the first three letters of the Greek alphabet: alpha, beta, gamma.

Their analysis, it turned out, also conflicted with known rates of nuclear reactions from weapons research during the second world war. A few months later, Gamow redid the analysis, removing the conflict by adding in the sea of thermal radiation. He realised that the radiation would still be present, but he was silent about how hot it might be. Alpher, with Robert Herman, supplied a figure, arriving at a temperature of about 5 kelvin.

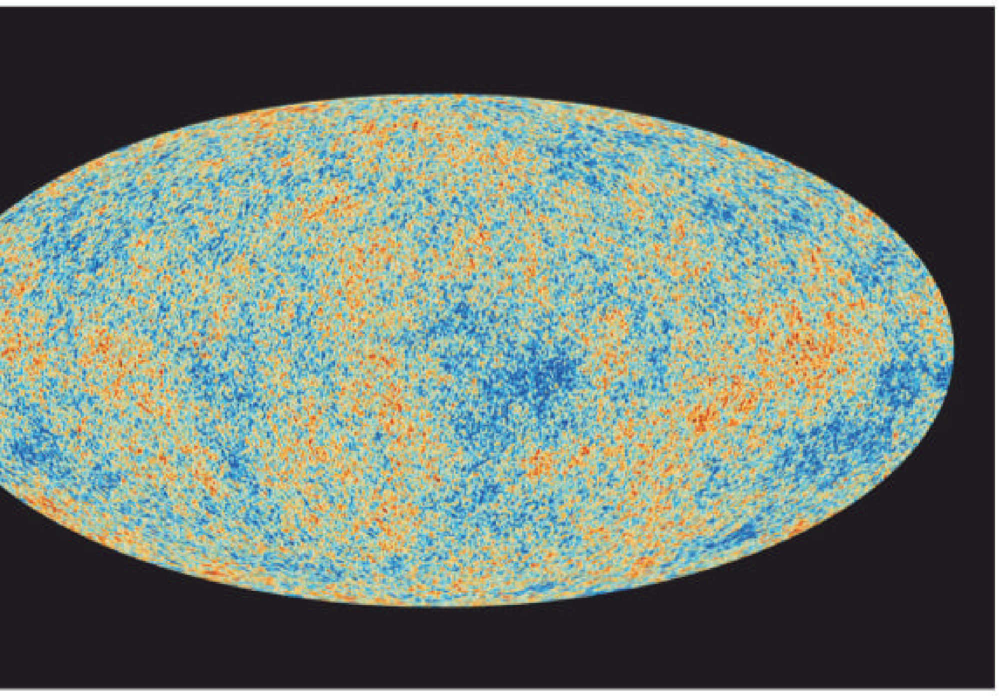
Gamow's analysis incurred Hoyle's displeasure, both because it ignored the steady state cosmology and because it produced an abundance of helium of about 30 per cent of the total cosmic mass, far higher than observations could at the time support. When, by the early 1960s, improved evidence suggested this figure is in fact about right, Hoyle was one of the first to note that it fits the big bang cosmology – but he took



NASA/COBE



NASA/WMAP



ESA/PLANCK COLLABORATION

care to suggest that it might instead have come from explosions of supermassive stars, little bangs in a steady-state universe.

This is kind of where we were in 1964. Bob was taken by the idea that an expanding big-bang universe might have bounced back after a previous cycle of expansion had collapsed. During such a collapse, starlight would be compressed along with the matter, reaching temperatures and densities high enough to pull apart the heavy elements produced by stars in the last cycle and so provide new nuclear fuel for our cycle. This process would cause the radiation to reach thermodynamic equilibrium, with the same temperature everywhere, producing a characteristic spectrum of intensities at different wavelengths known as a thermal Planck spectrum – a sure sign of our universe's origin in a hot, dense state.

Bob also knew microwaves. As part of the effort to develop shorter-wavelength radar with better resolution during the second world war, he had invented a microwave detector that was used to study radiation absorption and emission by atmospheric water vapour. As a sideline, he had written a paper in 1946 on the upper limits of cosmic microwave radiation. The experiments were done in Florida, whose muggy nights supplied ample atmospheric water to flood the detector with microwaves. Had they happened on a cold, dry winter's night in Princeton, Bob might have been the first to detect the CMB.

But people can be forgetful, and in 1964 we had to remind him that he had even set a limit. He also told us he had vague memories of hearing a lecture by Gamow, though not what it was about. But perhaps echoes of Gamow's thoughts, along with the Florida tests, were driving Bob's thinking in 1964. In the 1981 *New Scientist* article, Hoyle recalls a conversation he had with Bob in 1961 about the behaviour of interstellar CN, a molecule that rotates as if it were immersed in microwaves at › about 2.3 kelvin. Bob certainly didn't remember that discussion, for it would have been quite out of character for him not to tell us, but this indication that the CMB might exist may also have been lurking in the back of his mind.

I suppose "back of the mind" is as good a way as any to account for the sudden disruptions of orderly research by surprising and exciting ideas. That unexpected, even unwanted, measurements – an annoying hiss in a glorified telecommunications antenna – on occasion prove to be wonderfully informative is just one of the ways science advances.

I don't remember any expression of regret by Bob or any of us at being scooped. Instead, there was excitement that something was there to be measured and analysed. Early on, our focus was on finding out whether the CMB did indeed have that telltale Planck spectrum. In the 1970s, this was still sufficiently disputed to support the differing interpretations put forth in Rees's and Hoyle's articles.

Roll and Wilkinson's experiment soon added another data point, and Wilkinson eventually added many more, culminating in his leading role in the COBE satellite mission that, in the early 1990s, finally showed the CMB spectrum is close to a Planck form with a temperature of around 2.73 kelvin. (Herb Gush at the University of British Columbia, Canada, whose rocket-based experiment confirmed it one month later, might have scooped this finding, had it not been for launch delays.)

Few apart from Hoyle and his close associates doubted the universe's origin in a big bang by then. Penzias and Wilson had received a share of a Nobel prize in 1978, which is fair enough: they tracked down every conceivable terrestrial source of excess microwave radiation, and complained about their inability to account for the anomaly until someone at last paid attention.

But Bob ought to have shared it, both for inventing much of the technology used to discover the CMB and

THE COSMIC MICROWAVE BACKGROUND

For 380,000 years after the big bang, there was nothing but a ludicrously hot, dense soup of subatomic particles. Photons of light pinballed between them, unable to escape, trapped just as light is trapped in a dense fog.

At some point in the universe's expansion and cooling, however, things had cooled enough for hydrogen atoms to form. This cleared the subatomic obstacles from the way of the photons. It was as if someone had flicked a switch: light shot out in every direction and the cosmos became transparent. The cosmic microwave background is made up of those first liberated photons.

Those photons have been travelling in all directions through the expanding cosmos ever since. You can't see this light with the naked eye because, as the universe expanded, it stretched out from visible wavelengths into microwaves, hence the name. Sitting in our particular spot some 14 billion years later, we see them as a background radiation that suffuses the sky at an almost uniform temperature of 2.7 kelvin.

This light is very special, says Jacques Delabrouille at Paris Diderot University in France, and a rich source of information on the cosmos. "It is a backlight that shines on all of the structures between its emission and us, so it probes basically all the history of the universe."

Its general uniformity reinforces a simplifying assumption central to our attempts to extract a workable model of the universe from the complex equations of general relativity – the "cosmological principle" that the cosmos is pretty much the same everywhere. Meanwhile, tiny fluctuations in its temperature give clues as to the make-up of the universe – and support the idea that it is dominated by entities, dark matter and dark energy, whose identities we struggle to pinpoint.

for proposing the search that led the Bell researchers to recognise they had already found it.

Meanwhile, my own theoretical thinking started down the road that Gamow pioneered, but I kept seeing new things to do. The CMB would be disturbed by the gravitational attraction of matter, which in our universe is now quite clumped up, and by interactions with matter in the form of plasma in the hot young universe. More accurate measurements of how much the CMB departs from uniformity, from Wilkinson and a growing list of colleagues, drove me to devise the now established lambda-CDM cosmological model.

At the time, I thought I was writing down a theory that involved disturbances to the microwave background too small to be detected. Thanks to COBE and its follow-up missions, however – NASA's WMAP, the European Space Agency's Planck probe, and a host of others – we now have precise measurements of how the CMB departs from an exactly uniform sea of radiation that inform us about the history of expansion of the universe and the nature of its material contents. I have written three books on the subject, each much larger than the last. Now the field has grown much too big for a monograph; it is richer by far than anything I imagined when I started following Bob's suggestion a half-century ago.

With lambda-CDM, we now have an excellent fit of cosmological theory and measurements, albeit one that requires two hypothetical components: unseen cold dark matter – the CDM – to keep galaxies clumped together, and the cosmological constant, lambda. This constant is the quantity Einstein introduced into his equations of relativity to create a static universe, and then regretted as an inelegant and unneeded complication. Now it is needed to account for the accelerated expansion of the universe revealed by measurements of far-off supernovae, and also for details of the distribution of the CMB. Its new name, "dark energy", isn't a sign of progress: we still don't understand its nature.

→
Chapter 3 has more on dark matter and energy

Not least because of these two hypothetical interlopers, caution is in order. The fit of the lambda-CDM model also depends on an optimistic extrapolation of general relativity from the largest tested scale of the solar system to the vastly bigger scale of the observable universe. But tests from the CMB and elsewhere are abundant enough now that I am forced to conclude that we have a convincing approximation to what happened as the universe expanded and cooled.

Hoyle's steady-state cosmology is convincingly ruled out, although its philosophy reappears in the multiverse – the idea that universes like the one we live in are constantly budding out of some greater universe. The theory of eternal inflation, where this idea arises, postulates that our universe underwent a period of enormously accelerated expansion in its earliest instants, and offers an elegant way to understand what happened before the expanding universe had cooled to the point that it can be described by general relativity.

↓
See page 30 for more on inflation and the multiverse

The patterns by which the CMB's temperature deviates from uniformity over different distance scales matches what might be expected in a simple mathematical formulation of inflation, although the fit isn't unambiguous.

However that story pans out, cosmology has matured beyond all recognition in little over a century. In 1914, Einstein was putting the finishing touches to general relativity, the theory on which it is all based. Fifty years on from that, a crucial way station was reached: the identification of an unexpected hiss that tells the story of the universe's origin. ▮

WHAT EXACTLY WAS THE BIG BANG?

The basic premise of the big bang theory is clear: that the universe began in an infinitesimal pinprick of unimaginable heat and density that has slowly stretched and cooled into the cosmos we know today. How we interpret this "singularity", which seems to mark the very beginning of space and time, is another matter entirely.

A CENTURY ago, if you asked a cosmologist the universe's age, the answer may have been "infinite". These days, most agree that it has a finite age of about 13.8 billion years, give or take.

We can cross-check that against the oldest star we know of. HD 140283, aka the Methuselah star, is made almost entirely of hydrogen and helium, the predominant elements in existence in the big bang's aftermath. Astronomers reckon it is 14.46 billion years old, give or take 0.8 billion years. That could make it slightly older than the universe, but the fact that the age of the oldest star is so close to our estimates of the universe's age suggests we are on the right general track.

Getting your head around what an expanding universe of finite age implies is significantly less easy, even for seasoned practitioners. For cosmologist Martin Rees at the University of Cambridge, there are two strategies: bury yourself in equations, or draw pictures. "I'd put myself in the picture camp," he says.

Rees's trick is to imagine himself at one node of a

> “The big bang marks the point where the equations of general relativity break down”

three-dimensional lattice stretching as far as the mind’s eye can see, with the nodes linked by rods, all of which are expanding. That way you can visualise the universe moving away from you in all directions – while recognising that you would see the same thing from any other node. “You understand there is no central position,” he says. Equally, if you wonder where the big bang occurred, you can wind this picture backwards, with the rods all contracting together until the nodes are all in the same place – and you recognise that, wherever you are in the cosmos, the big bang happened here, there and everywhere.

Using our standard model of cosmology, we can also make a reasonable stab at assessing how big the universe has got. The most distant galaxy known is GN-z11. Light from it has taken 13.4 billion years to reach us, most of the age of the universe. Working from the rate of expansion given by the standard model, this galaxy is probably now about 32 billion light years away from us. Extrapolating to the entire observable universe, astronomers estimate it has a diameter of 93 billion light years, or roughly 10^{26} metres (100 million billion billion kilometres).

But that is just the distance between the furthest things we can see: the universe has no discernible edge. “You don’t walk 10^{26} metres and then hit a brick wall,” says Tony Padilla at the University of Nottingham, UK. “The universe goes beyond that.” Periods of accelerated expansion thought to have happened during the universe’s earliest instants (thanks to cosmic inflation) and in more recent aeons (thanks to dark energy) mean that the horizon of the observable universe is by no means the end of all things. Beyond it are galaxies we will never see because the intervening space is expanding too fast, so their light can never reach us.

↓ Read on to the next section for more on inflation

→ Turn to page 43 for more on dark energy

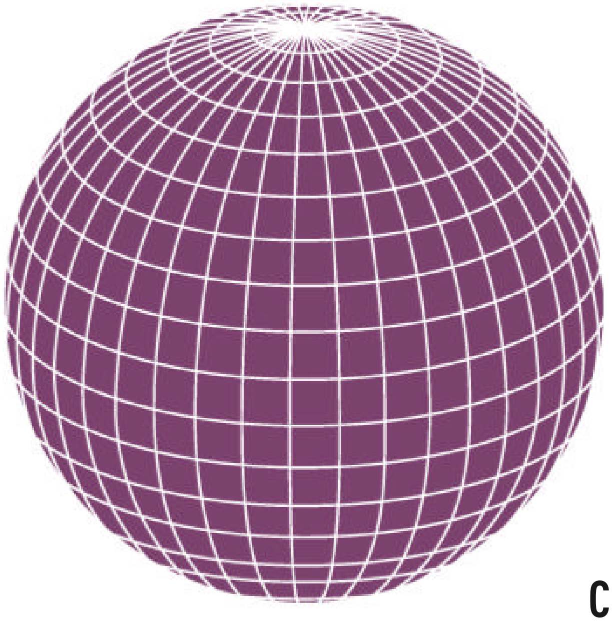
The most challenging thing of all, however, is again to wind back and envisage what came before the big bang – or what the moment itself signifies. Our current conceptions of physics suggest that the question makes no sense: as we rewind time to the very first instants, the intense concentration of energy jumbles up even space and time in a confusion of stuff. “There is no direction of time, so there’s no before and after,” says Rees. “The analogy that’s always made is it’s like asking what’s north of the North Pole.”

In terms of the mathematics of general relativity, the moment of the big bang isn’t an event, but a “singularity” – a point where quantities become infinite and the theory simply breaks down. For most cosmologists, a true understanding of what it means requires a theory that goes beyond general relativity and unifies it with quantum theory, the other great pillar of modern physics that explains the workings of all the other forces of nature bar gravity. ▮

→ Page 88 has more on quantum big bangs

GARRY KILLIAN/ISTOCK

INFLATION AND THE MULTIVERSE


CLOSED UNIVERSE

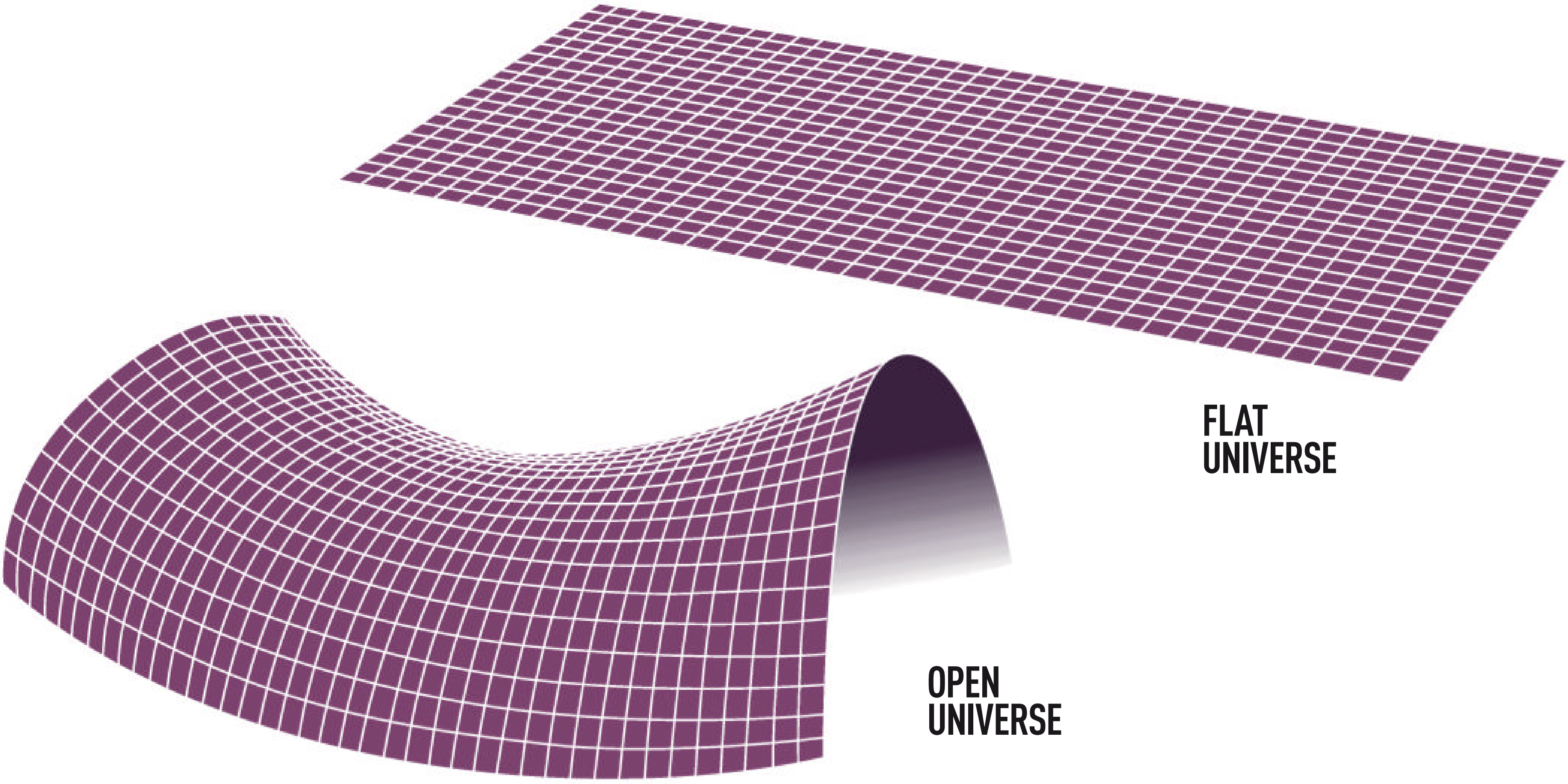
As we have seen, the existence of the cosmic microwave background is a spectacular confirmation of the big bang theory. But lurking within that success are some problematic details. Attempts to explain them away have led cosmologists to some truly out-of-this-world conclusions.

SEEN from Earth, the cosmic microwave background is pleasingly uniform in temperature in all directions in the sky. Measure it 10 billion light years away in one direction and 10 billion light years in the other, and you still observe that pleasing uniformity.

That is where the problems start. Run a simple story of a steadily expanding cosmos backwards, and it would take 20 billion years for these patches of space to meet – more than the age of the universe. In a simple big-bang universe, they were never close enough to equalise their temperatures. The uniformity of the cosmic microwave background becomes a highly improbable coincidence.

It isn't the only issue. In our expanding universe, there are basically two possibilities for the overall geometry of space-time. If the gravity produced by all matter is stronger than the expansion, it will ultimately pull everything back together. In that case, we are living in a "closed" or spherical universe. If whatever is driving the expansion overpowers gravity, however, then we have a perpetually expanding or "open" universe that looks like a saddle (see diagrams, above).

Yet information encoded in the cosmic background shows space is extremely flat: it is neither open



The cosmos is most likely to be closed and positively curved (left) or open and negatively curved (above). And yet it seems to be almost exactly flat (top) – a coincidence that is hard to explain

nor closed, but Euclidean geometry reigns supreme and parallel lines never meet. Like the uniformity, this flatness is also highly unlikely, given what we know about gravity and its compulsion to warp space. Again, a simple big bang cannot explain it.

This is the background to a theory that, when it was proposed in 1980 by Alan Guth, then a young postdoctoral researcher at Cornell University in Ithaca, New York, came as a godsend. It is known as cosmic inflation.

At the beginning of time, so the idea goes, all that existed was a quantum field associated with a hypothetical particle called the inflaton. It found itself in a "false vacuum" – a state that is temporarily stable, but not its lowest, true-vacuum energy state. It is as if the inflaton were poised on a small plateau on a steep mountainside. All things being equal, it could rest there undisturbed, but the slightest jiggle would send it careering down towards the true vacuum below.

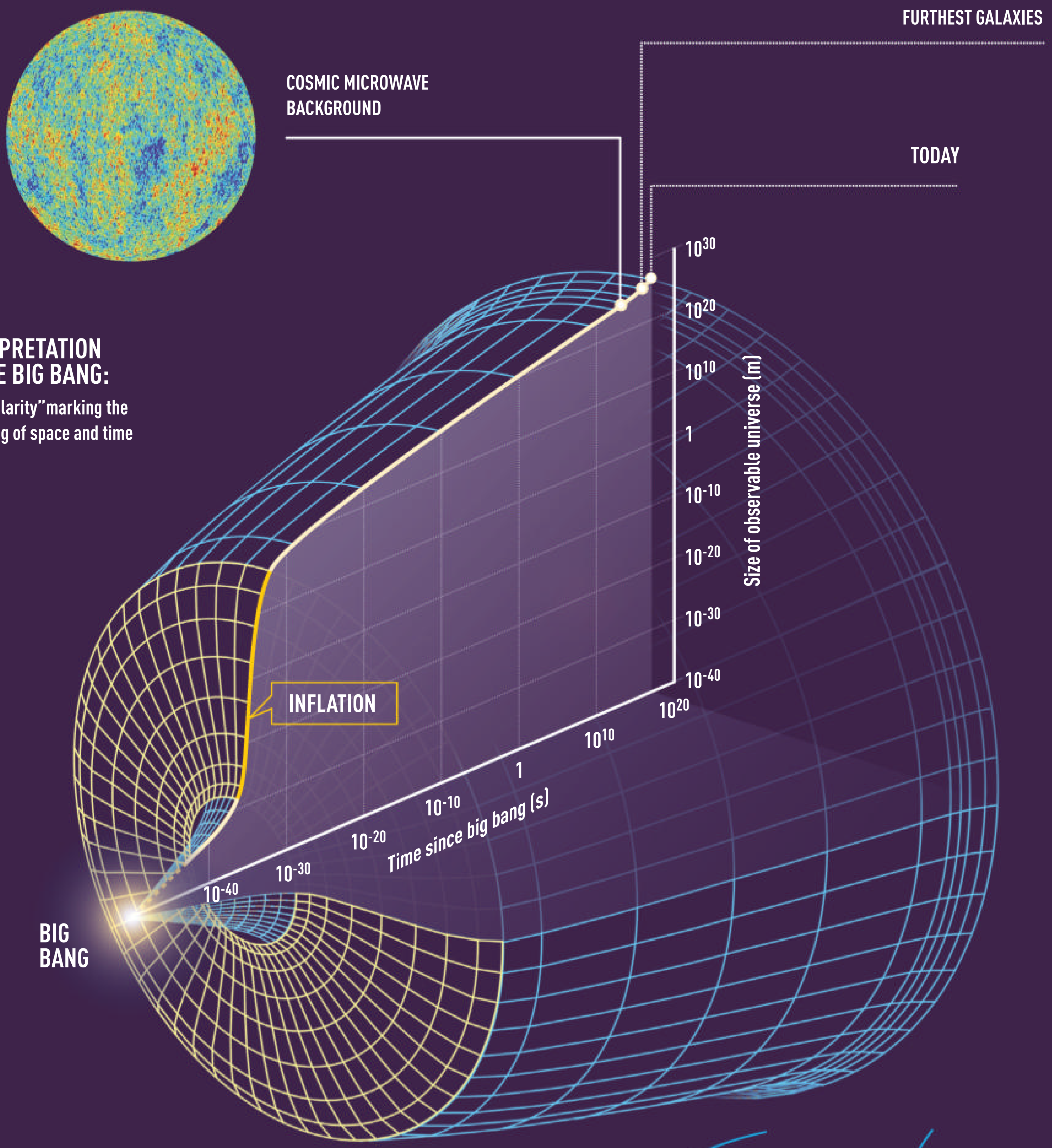
In the universe's first fraction of a second, all things were not equal. Random quantum fluctuations in energy provided just the jiggle to set the inflaton on its way. As it fell towards the true vacuum, it generated a kind of repulsive gravity that pushed the space out around it. The further it fell, the more it pushed until space was ballooning outward at a speed far greater than that of light.

This is physically all above board. Relativity forbids objects from travelling faster than light through space, but places no constraints on what space itself can do. And when the inflaton hit rock bottom, all the kinetic energy it acquired in its headlong descent poured into the universe, creating not just the radiation that fills it, but – thanks to that bedrock equation of special relativity, $E = mc^2$ – the matter that went on to form stars, planets and, eventually, us.

All this happened in considerably less than the blink of an eye: in just 10^{-33} seconds, the observable universe ballooned over 20 orders of magnitude in size, from a diameter about a billionth that of an atomic nucleus to a mini-cosmos about a centimetre across (see diagram, overleaf).

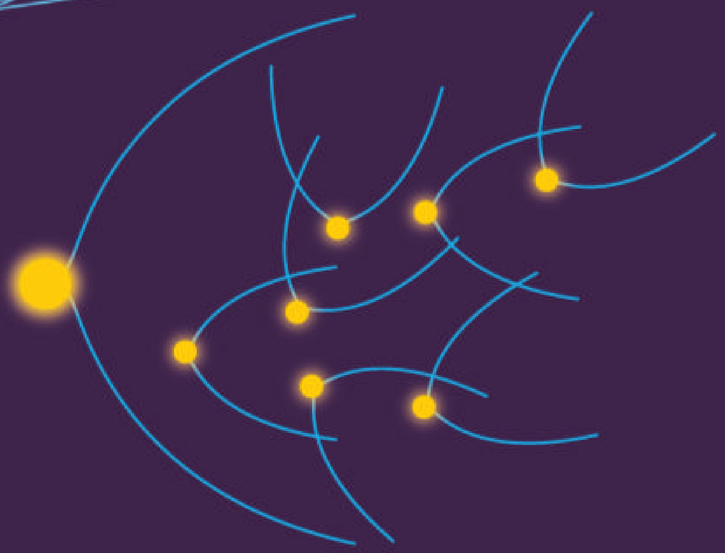
In one fell swoop, inflation solved the big bang's problems. Those patches of sky no longer need a 20-billion-year rewind to have met and mingled: ›

By rapidly pushing apart the early universe, a period of inflation can explain why distant parts of the cosmic microwave background look like they came from the same place



INTERPRETATION OF THE BIG BANG:

A "singularity" marking the beginning of space and time



PROBLEM: Once inflation starts, it cannot stop

Bits of the inflating universe themselves begin inflating off into independent existences. This creates an infinite "multiverse" of universes, making cosmological predictions impossible

SOURCE: WMAP/NASA

inflation gave them the shove-off they needed to ensure they arrived far faster at the far reaches of the cosmos. And that absurdly unlikely flatness is nothing of the sort: inflation makes the universe so large that any measurable region must look flat, the same way that the ground at your feet looks flat even though Earth's surface is curved.

Inflation's munificence didn't stop there. By inflating tiny quantum fluctuations in the density of the cosmos to astronomical proportions, it produced a blueprint for how stuff clumped into ever-larger agglomerations of matter such as the galaxies we see today. All of these successes mean inflation has become a canonical part of the story of the big bang. Without it, we simply can't explain the universe we see.

"That would have been the perfect point for inflation to bow, wait for applause and exit stage left," says cosmologist Max Tegmark at the Massachusetts Institute of Technology. But that didn't happen. Instead, inflation kept on predicting still more things. Things that nobody wanted – like other universes.

Once inflation starts, it is nearly impossible to stop. Even in the tiny pre-inflation cosmos, quantum fluctuations ensured that the inflaton field had different energies in different places – a bit like a mountain having many balls balanced precariously at different heights. As each one starts rolling, it kicks off the inflation of a different region of space, which races away from the others at speeds above that of light. Because no influence may travel faster than light, these mini-universes become completely detached from one another. As the inflaton continues its headlong descent in each one, more and more bits of space begin to bud off to independent existences: an infinite "multiverse" of universes is formed.

This isn't good news for our hopes for cosmic enlightenment. In a single universe, an underlying theory of physics might offer a prediction for how flat the universe should be, say, or for the value of dark energy, the mysterious entity that seems to be driving an accelerated expansion of the universe. Astronomers could then go out and test that prediction against observations.

→ Chapter 3 has more on dark energy

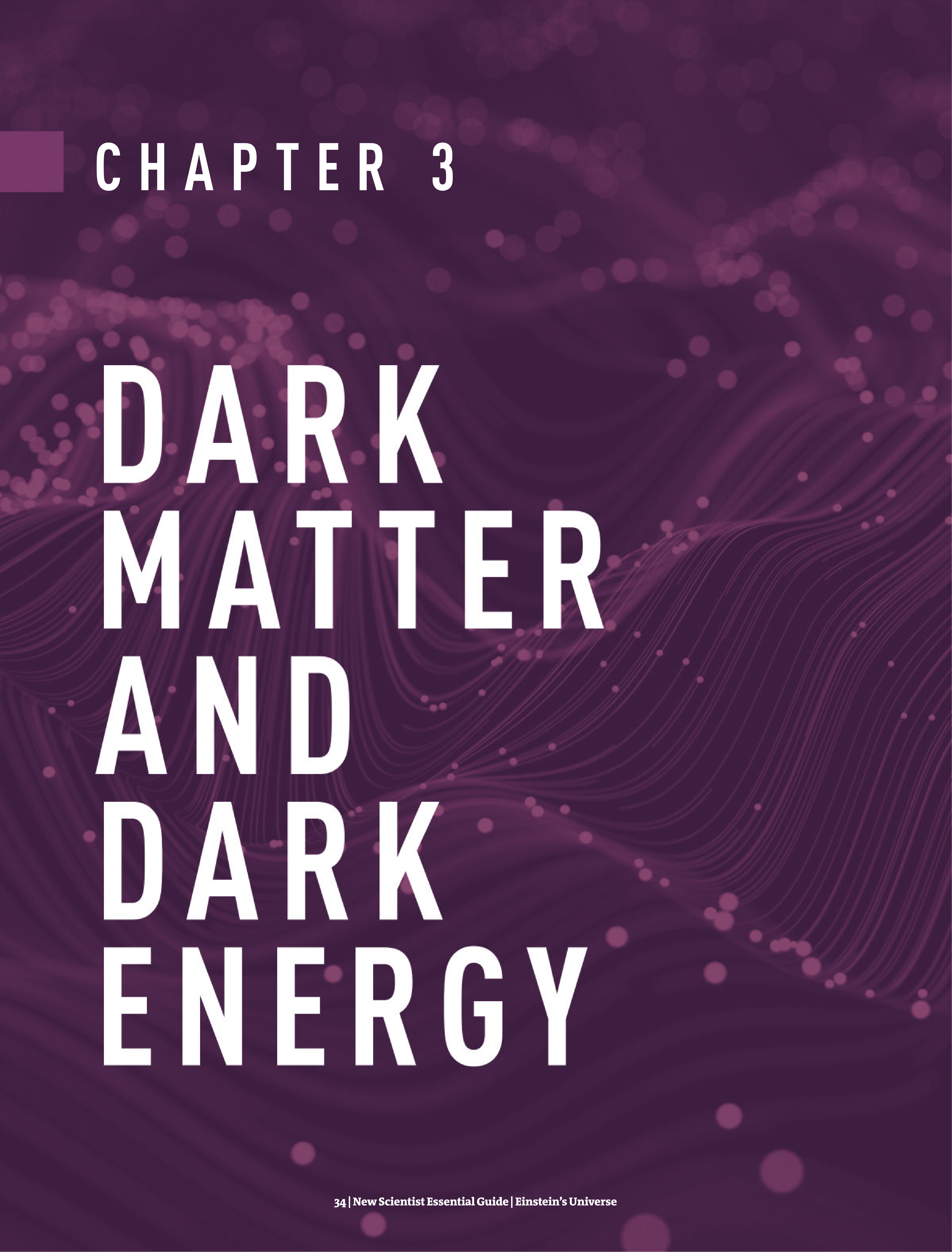
That isn't possible in an infinite multiverse: there are no definite predictions, only probabilities. Every conceivable value of dark energy or anything else will exist an infinite number of times among the infinite number of universes, and any universal theory of physics valid throughout the multiverse must reproduce all those values. That makes the odds of observing any particular value infinity divided by infinity: a nonsense that mathematicians call "undefined".

At first, cosmologists hoped to make sense of these infinities by taking a finite snapshot of the multiverse at some particular time, and then extrapolating the relative probabilities of various observations out to later and later times and an ever larger number of universes. But relativity stymies that approach. It means there is no single clock ticking away the seconds of the multiverse, and there is an infinite number of ways to take snapshots of it, each giving a different set of probabilities. This "measure problem" destroys inflation's ability to make predictions about anything at all, including the smoothness of the cosmic background, the curvature of space or anything else that made us believe in the theory in the first place.

"We thought that inflation predicted a smooth, flat universe," says Paul Steinhardt at Princeton University, a pioneer of inflation who has become a vocal detractor. "Instead, it predicts every possibility an infinite number of times. We're back to square one." Tegmark agrees: "Inflation has destroyed itself. It logically self-destructed."

Sean Carroll, a cosmologist at the California Institute of Technology, is more circumspect. "Inflation is still the dominant paradigm," he says. "But we've become a lot less convinced that it's obviously true." That isn't just because of the measure problem, he says. More basically, we don't know what an inflaton field is, why it was in a false vacuum and where it and its energy came from. Having an inflaton field perched so perfectly and precariously atop that mountainside seems no more likely than the flukes the idea was intended to explain. "If you pick a universe out of a hat, it's not going to be one that starts with inflation," says Carroll. ∎

→ See page 88 for more on alternatives to inflation



CHAPTER 3

DARK MATTER AND DARK ENERGY

To cosmologists, the stars that define our view of the universe from Earth are an insignificant decoration on the true face of space. Far outweighing ordinary stars and gas are two elusive entities, dark matter and dark energy. Together, they make up more than 95 per cent of all the stuff in the universe.

Both are needed to make the standard model of cosmology based on general relativity work. Dark matter supplies enough gravity to stop galaxies disintegrating as they rotate, while dark energy explains why, contrary to our naive expectations, the universe's expansion appears to be accelerating.

With dark matter, we at least have a clue what we are looking for; with dark energy, we are struggling even to define what the nature of our quarry is. But do these spectres haunting cosmology foreshadow a deeper problem with our theories?

THE STANDARD COSMOLOGICAL MODEL

Our picture of the universe based on general relativity is supremely successful, but perhaps only because, with its introduction of dark matter and dark energy, most of it is made up. Are these weighty phantoms too great a burden for our observations to bear – a wholesale return of conjecture out of a trifling investment of fact, as Mark Twain put it?

PREVIOUS PAGE: JUST_SUPER/ISTOCK
PHILIPP TUR/ISTOCK

THE ghosts first entered the cosmological machine back in the 1930s, although no one took much note back then. That was when astronomer Fritz Zwicky noticed that galaxies towards the edge of the Coma cluster of galaxies were rotating faster around the cluster's centre than they should have been, given the amount of visible matter there was.

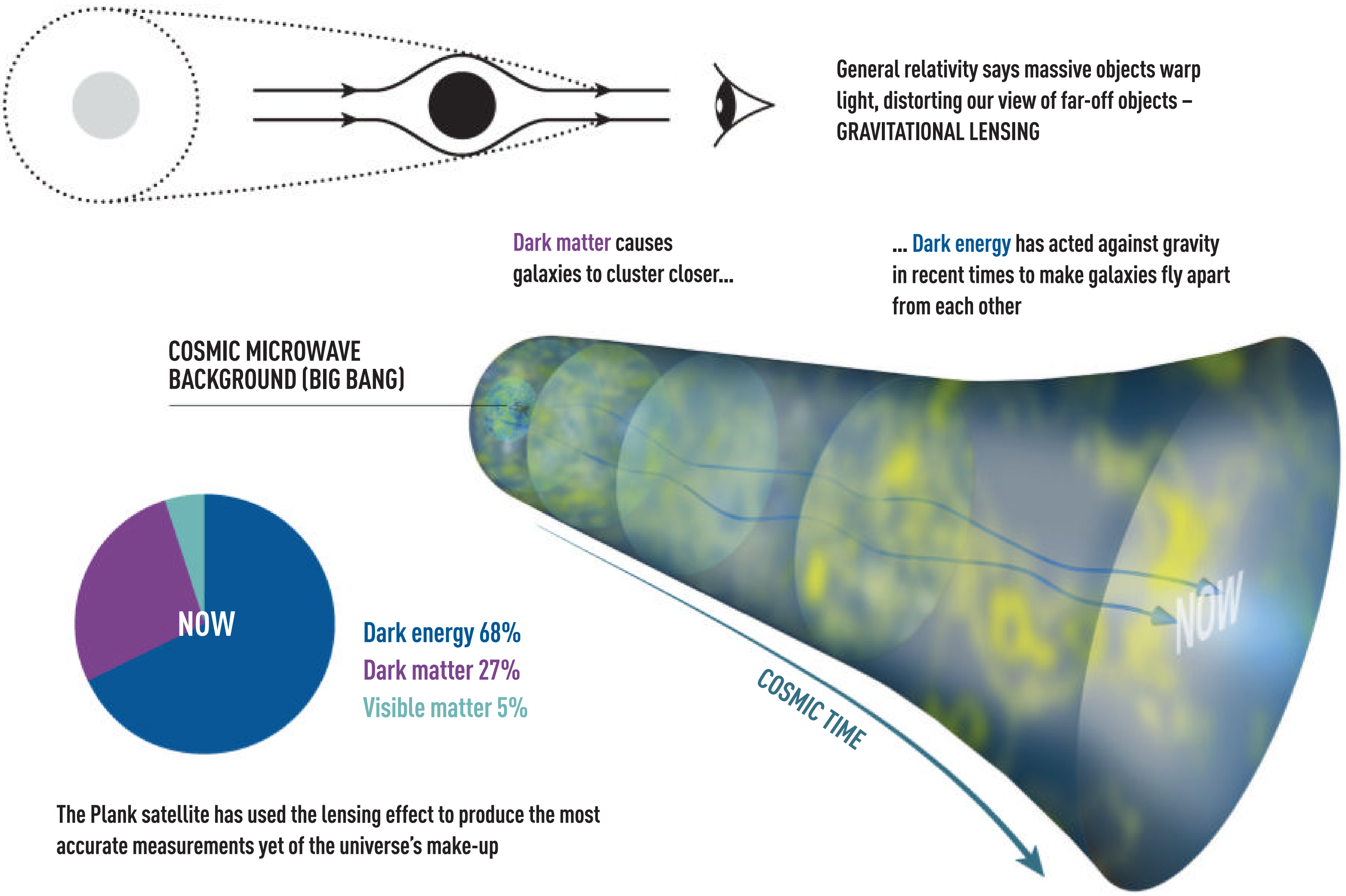
But the idea only really gained traction in the late 1970s, when astronomers Vera Rubin and Kent Ford found a consistent effect in rotating spiral galaxies: stars were orbiting their galactic centres faster than we would expect based on how massive the galaxies are presumed to be, just counting stars and adding their masses together. The gravity of the visible matter would be too weak to hold these galaxies together according to general relativity, or indeed plain old Newtonian physics.

There are two ways to address this inconsistency: there is either more matter than we can see in these galaxies, or our theory of gravity is wrong. That second possibility is the basis of modified gravity theories that can't entirely be discounted, but have difficulty reproducing other aspects of the universe as we see it (see "Have we got gravity wrong?", page 42).

The existence of some form of dark matter, meanwhile, has since been backed up by other lines of evidence, such as how groups of galaxies move, and the way they bend light on its way to us. Overall, there appears to be about five times as much dark matter as visible gas and stars. Its identity is unknown, but it seems to be something beyond the standard model of particle physics. Despite our best efforts, we have

›

Photons from the cosmic microwave background left over from the big bang get distorted on their way to us. Measuring those distortions tells us how the balance of dark matter and dark energy has changed all the universe's life


General relativity says massive objects warp light, distorting our view of far-off objects – GRAVITATIONAL LENSING

Dark matter causes galaxies to cluster closer…

… Dark energy has acted against gravity in recent times to make galaxies fly apart from each other

The Plank satellite has used the lensing effect to produce the most accurate measurements yet of the universe's make-up

yet to see or create a dark matter particle on Earth.

For all that, though, dark matter changed cosmology's standard model only slightly: its gravitational effect in general relativity is identical to that of ordinary matter, and even such an abundance of gravitating stuff is too little to halt the universe's expansion.

The second dark spectre required a more profound change. In the 1990s, astronomers traced the expansion of the universe more precisely than ever before, using measurements of explosions called type 1a supernovae. They showed that the cosmic expansion is accelerating. It seems some repulsive force, acting throughout the universe, is now comprehensively trouncing matter's attractive gravity: dark energy.

Observations of the distribution of matter in the universe, and tiny fluctuations in the cosmic microwave background (see diagram, above), leave us with a precise recipe for the cosmos. The average density of ordinary matter in space is 0.426 yoctograms per cubic metre (a yoctogram is 10^{-24} grams, and 0.426 of one equates to about 250 protons), making up less than 5 per cent of the total energy density of the universe. Dark matter makes up 27 per cent, and dark energy 68 per cent. The resulting standard model of cosmology is often known as lambda-CDM, in tribute to its main components: the Greek letter lambda denotes the cosmological constant; the CDM is "cold dark matter".

So our model of a big-bang universe based on general relativity fits our observations very nicely – as long as we are happy to make 95 per cent of it up. "We don't know what dark energy is and we don't know what dark matter is, and that should be a little bit embarrassing," says Robert Kirshner, a cosmologist at Harvard University and a member of one of the supernova teams that first exposed dark energy.

Kirshner sees that as a challenge. "It doesn't mean there is any flaw in our arguments. It gives a sense not of desperation, but inspiration." But as long as we have no evidence of dark matter in the lab or a proven physical basis for dark energy, the possibility remains that we are living under some profound misapprehension – an unknown unknown, something so basic awry in our mathematical model of the universe that, as yet, we haven't been able to imagine the form of our mistake. ▮

Dark matter's no-show might indicate that our models of the early universe are missing a crucial piece, says Dan Hooper.

THE MYSTERY OF DARK MATTER

PROFILE
DAN HOOPER

Dan Hooper is a cosmologist and head of theoretical astrophysics at Fermilab in Batavia, Illinois. He is the author of books including *At the Edge of Time: Exploring the mysteries of our universe's first seconds*

WE SEE its effects in how stars move within galaxies, and how galaxies move within galaxy clusters. Without it, we can't explain how such large collections of matter came to exist, and certainly not how they hang together today. But what it is, we don't know.

Welcome to one of the biggest mysteries in the universe: what makes up most of it. Our best measurements indicate that some 85 per cent of all matter in our universe consists of "dark matter" made of something that isn't atoms. Huge underground experiments built to catch glimpses of dark matter particles as they pass through Earth have seen nothing. We hoped that particle-smashing experiments at the Large Hadron Collider (LHC), buried underground near Geneva in Switzerland, would create dark matter – but they haven't, at least as far as we can tell. The hunt for dark matter was never supposed to be easy. But we didn't expect it to be this hard.

Let's start with what we do know about this substance – or perhaps substances. Dark matter isn't familiar atomic matter or any of the exotic forms of matter created at the LHC or at other particle accelerators. It doesn't appreciably interact with itself or with ordinary matter, except via gravity. It can pass through solid objects like a ghost and doesn't emit, absorb or reflect any easily measurable quantities of light. It is invisible, or at least nearly so.

Yet without dark matter, it is unlikely that we would be here. As galaxies and galaxy clusters were built up, dark matter played the role of scaffolding: it gathered into enormous clouds whose gravity attracted and pulled together the atomic matter that would ultimately form the luminous bit of galaxies. Without the gravity of dark matter holding stars in place, they would fly outwards, in some cases escaping into intergalactic space. Many galaxies would simply disintegrate.

›

We see dark matter's imprint in many other ways, too, for example in how a galaxy cluster's gravity deflects light that passes it. Perhaps the best evidence of all for dark matter's existence comes from temperature patterns observed in the cosmic microwave background, the radiation left over from the big bang. Measurements of this radiation provide us with a map of how matter was distributed throughout our universe only a few hundred thousand years after its beginning. This map tells us that our universe was very uniform in its youth, with only the smallest variations in density. Without help from dark matter, there is no way that these density variations could have grown fast enough to form the galaxies and other large structures of today's universe.

←
Turn back to page 23 for more on the cosmic microwave background

A decade or more ago, many physicists, including me, thought we knew what dark matter was likely to consist of: weakly interacting massive particles, or WIMPs. As their name suggests, these are relatively heavy particles that, besides gravity, only interact via the weak nuclear force, which also governs subatomic processes such as radioactive beta decay. WIMPs seemed compelling because we could understand how they would have been created in the early universe.

During the first millionth of a second or so after the big bang, all of space was filled with a hot, dense plasma in which all sorts of particles, from photons and electrons to top quarks and Higgs bosons, were constantly created and destroyed. As space expands, however, the temperature of the plasma steadily drops. Eventually, it can't supply the energy required to make heavier particles, and their production stops. When this happens to a species of particle, most are destroyed – annihilated – and converted into other forms of energy. How many survive depends on how the particles interact and how often.

This leads us to a happy coincidence: for a particle species to emerge from the big bang with an abundance equal to that of dark matter today, it must have interacted through a force about as powerful as the weak nuclear force. A stronger force would have caused too many particles to be destroyed, while a feebler force would have allowed too many to survive. Rather like the temperature of Goldilocks's porridge, the strength of the weak force seems just right to explain how dark matter came to be formed.

But that story now seems rather like a fairy tale. If dark matter does consist of WIMPs, we can estimate how much it should interact with ordinary atomic matter via the weak force, and so design experiments to detect it. These experiments, housed in deep underground laboratories to avoid the constant bombardment of cosmic radiation, started out small,

SEVEN WAYS TO MAKE DARK MATTER

WIMPS

The textbook solution to dark matter is that it is a thick, slow-moving soup of weakly interacting massive particles (WIMPs). That could explain the odd way galaxies rotate – yet no detector has yet found a WIMP. If they do exist, it seems they must be lighter than we thought.

MACHOS

This is the idea is that dark matter is just normal stuff hiding at the edges of galaxies, "massive astrophysical compact halo objects" that are so dim as to be invisible. Candidates include black holes or failed stars. Alas, MACHOs could only account for a tiny fraction of the universe's missing mass.

MACROS

It could be that the dark stuff is made of dense clumps of quarks, the particles that, in pairs or triplets, form ordinary matter. These "macros" could be as dense as neutron stars and extremely heavy. Unfortunately, the experiments needed to spot them, such as deploying seismometers on the moon, are too outlandish to carry out.

deploying detectors of only a few kilograms of crystalline materials such as germanium, calcium tungstate or sodium iodide, sensitive to the light, heat and electric charge that would be produced in collisions of WIMPs with normal matter.

Over the past two decades, the size and sophistication of these experiments has hugely increased. The latest iterations are enormous, deploying anything up to tonnes of liquid xenon as their detectors. These experiments – XENON1T under the Gran Sasso mountain in Italy, LUX in South Dakota and PandaX-II in Sichuan, China – are each roughly 10,000 times as sensitive as the most sophisticated dark matter detectors operating in 2006.

But they, too, have failed to turn up WIMPs. The only experiment that even claims to have detected anything resembling dark matter goes by the name of DAMA. Most researchers think the signal it picked up is almost certainly produced by something else: a long list of other experiments have searched for the kinds of WIMPs that could have made it, but have seen nothing.

The only other possible piece of evidence we have for WIMPs comes in the form of a strange gamma-ray signal seen emanating from the centre of the Milky Way. My collaborators and I spotted this signal in data from NASA's Fermi space telescope more than a decade ago. It took years for us to convince most people that it was real. We continue to debate whether these gamma rays are produced by dark matter or by something else, such as a group of thousands of rapidly spinning neutron stars. At the moment, we just can't be sure.

Whatever the resolution of that argument, the longer we go without directly detecting WIMPs, the more we are forced to confront the uncomfortable possibility that they might not be there. And yet dark matter must exist – alternative explanations, such as modifying gravity to produce the same sort of effects, don't seem to work (see "Have we got gravity wrong?", overleaf). If not WIMPs, then what?

One possibility is that dark matter could interact with other forms of matter and energy even less than we had imagined – perhaps only through gravity or some force so feeble that we haven't even discovered it yet (see "Seven ways to make dark matter", below). Such a particle would be even more difficult to detect in underground experiments or to produce with particle accelerators. The problem is that such non-interacting particles would probably survive the big bang in vast numbers, and wildly exceed the abundance of dark matter in our universe today. But if they interact rarely enough, perhaps these particles were never produced in great quantities in the first place, instead building up an appreciable abundance only gradually over the first fraction of a second of cosmic history.

It could be, too, that dark matter is just one of several kinds of particles that almost never interact with any known forms of matter and energy. This "hidden sector" of particles would involve forces ›

AXIONS

A punier version of the WIMP, axions would interact even less with ordinary matter. That suggests WIMP detectors might have spotted them – but they haven't. The jury is still out, at least until dedicated experiments such as the Axion Dark Matter Experiment return a verdict.

STERILE NEUTRINOS

Neutrinos pass through other matter almost as if it doesn't exist, but they are too light and zippy to be dark matter. Sterile neutrinos are a heavier, more aloof version. Signs of them have emerged in underground detectors, only to quickly disappear. We have also seen a suggestive excess of X-rays coming from galaxy clusters, but have failed to pin down the sources.

GRAVITINOS

The graviton is a particle proposed by the theory of supersymmetry to mediate the force of gravity, and the gravitino is its hypothetical "superpartner". It nicely fits the bill for a dark matter particle. The trouble is that there is still no sign of the many heavy partner particles predicted by supersymmetry.

WHITE HOLES

Black holes in reverse could provide an identity for dark matter – if they exist. See page 66 for more on this possibility.

and interactions that we have never observed, and that allow dark matter to evolve in a rich variety of ways. These interactions may have depleted the amount of dark matter, without leading to any appreciable interactions with ordinary matter.

The hidden-sector particles might become bound to each other, forming dark nuclei or dark atoms. One day, we could even discover something like a periodic table of the hidden sector elements. For that reason, of all the plausible ideas about dark matter that have grown in popularity in recent years, this is perhaps my favourite.

Many of these alternative dark matter candidates call for experiments very unlike those designed to hunt for WIMPs. One example is the Axion Dark Matter Experiment based at the University of Washington in Seattle and managed by scientists at my institute, Fermilab. It uses powerful magnetic fields to try to convert one hypothesised type of ultra-light dark matter particle, axions, into photons.

There is an even more dramatic possibility that many cosmologists are considering. Our surprise at dark matter's no-show is based on our current understanding of the early universe. Maybe we haven't seen the particles because dark matter is different from what we had expected – or perhaps because the universe's first moments were.

The amount of dark matter that was created in the big bang and survived it depends on how our universe evolved during its hot and volatile youth. We know a great deal about most of our universe's 13.8-billion-year history, but we have no direct observations that enable us to study the first fraction of a second, the window in which dark matter is thought to have formed.

Perhaps an experimental breakthrough will change the game yet again. But the stubborn elusiveness of dark matter has left many physicists and cosmologists surprised and confused. In droves, we are returning to our chalkboards, revisiting and revising our assumptions – and, with bruised egos and a bit more humility, desperately attempting to find new ways to make sense of a very dark and hidden universe. ∎

HAVE WE GOT GRAVITY WRONG?

Despite dark matter's long-standing refusal to reveal itself, most physicists remain confident that it exists – the evidence in its favour is just too great. A few, however, champion a very different possibility. Rather than explaining the motions of stars around galaxies with new forms of matter, they speculate that a different conception of gravity may be the answer.

These ideas fall under the general umbrella of modified Newtonian dynamics, or MOND. This postulates that gravity works ordinarily here on Earth and in our solar system, but differently in the low-acceleration environments experienced by stars throughout the Milky Way and other galaxies.

In these circumstances, the force of gravity is effectively stronger than Newton or Einstein thought. This strengthening of gravity creates the illusion that unseen dark matter must be present.

Many versions of MOND have been proposed over the past few decades, but they have suffered from a range of problems, both observational and theoretical. Perhaps the single biggest failure is MOND's inability to explain the temperature patterns observed in the cosmic microwave background. Whereas dark matter enables us to explain and understand the observed features of this light in incredible detail, no version of MOND has ever remotely done the same. Compounding these problems is the fact that no version of MOND has yet been able to explain the observed dynamics of galaxy clusters.

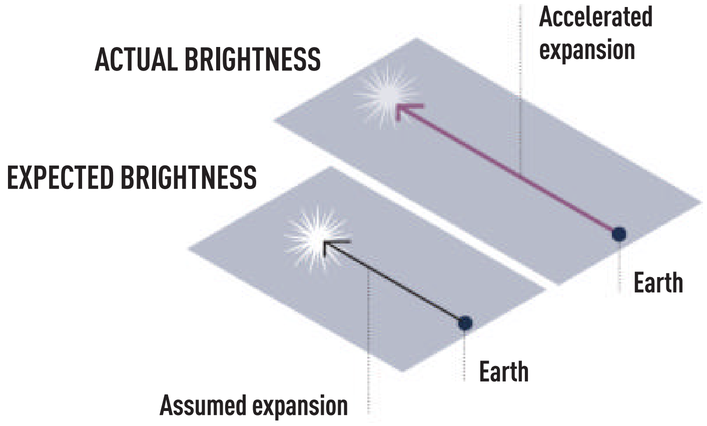
WHAT IS DARK ENERGY?

The discovery a quarter of a century ago that the universe's expansion is accelerating was a shocker. Is that an expression of a new field, a new force or just our own ignorance? It might make up more than two-thirds of the cosmos, but dark energy just keeps us guessing.

Many lines of evidence point to a mysterious dark energy countering gravity's pull and accelerating the universe's expansion

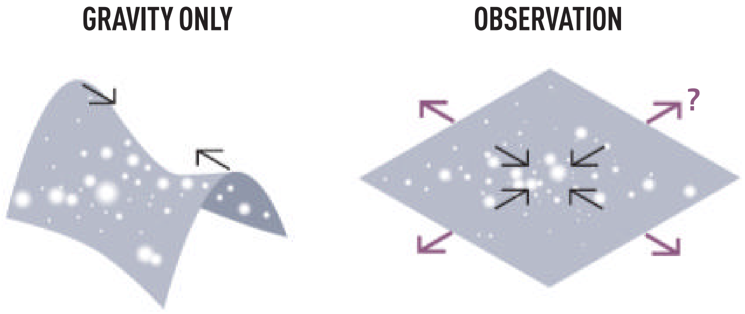
SUPERNOVAE

Distant type-1a supernovae are dimmer than expected - suggesting they are further away



COSMIC MICROWAVE BACKGROUND

With only matter's gravity at work, the shape of the universe should be curved. Patterns in the big bang's afterglow suggest it is nearly flat



OUR knowledge of dark energy's nature is limited to perhaps three things. First, it pushes. We first noted that in 1998, in the unexpected dimness of certain supernova explosions, which told us they were further away than we expected. Space seems to have begun expanding faster at some point, as if driven outwards by a repulsive force acting against the attractive gravity of matter.

Second, there is a lot of the stuff. The motion and clustering of galaxies tells us how much matter is abroad in the universe, while the cosmic microwave background radiation emitted 380,000 years after the big bang allows us to work out the total density of matter plus energy. This second number is much bigger. According to the latest data, including observations from the European Space Agency (ESA)'s Planck satellite, about 68 per cent of the universe is in some non-material, energetic, pushy form.

Third, dark energy makes excellent fuel for the creative minds of physicists. They see it in hundreds of different and fantastical forms.

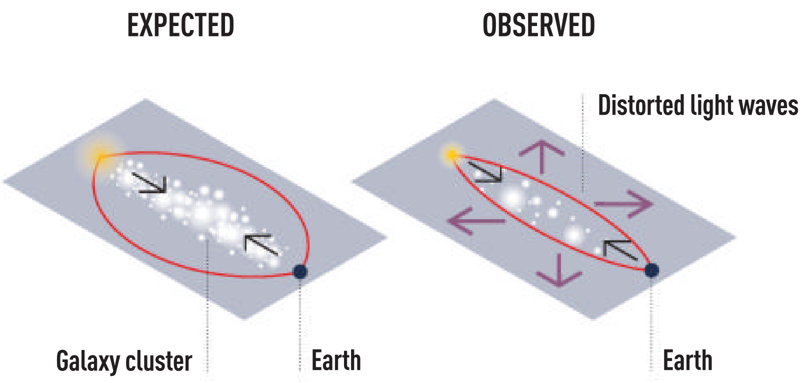
The tamest of these is the cosmological constant, the revival of Einstein's idea of an energy density inherent to space that creates a repulsive gravity – something, as we have seen, that he originally invented to create a static universe that neither expands or contracts. As space expands, there is more and more of the stuff, making its repulsion stronger relative to the fading gravity of the universe's increasingly scattered matter – explaining an important feature of dark energy, that its accelerating effects seem to have kicked in only in the past few billion years of the universe's history.

Particle physics even seems to provide an origin for this energy, in virtual particles that appear and disappear in the bubbling, uncertain quantum vacuum. The trouble is these particles have far too much energy – in the simplest calculation, about 10^{120} joules per cubic kilometre.

›

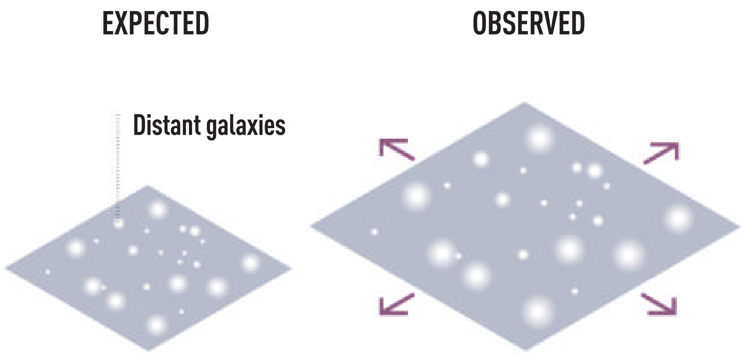
GRAVITATIONAL LENSING

Images of distant galaxies are distorted by intervening matter. The distortion observed indicates a repulsive force is stopping matter clumping



GALACTIC CLUMPING

The typical distances between galaxies far from Earth is different from what we expect, again suggesting that something is working against gravity's attractive pull



This catastrophic discrepancy leaves room for a menagerie of alternative ideas. Perhaps the most popular, first proposed in the 1980s, is to treat dark energy as an all-pervading field. Akin to a fifth force of nature, or "quintessence", its strength shifts over time. This relaxes the requirement that dark energy's density must remain constant over the lifetime of the universe. The acceleration attributed to the cosmological constant is itself increasing – and so the cosmological constant is anything but. Alternatively, it might be a modified form of gravity that repels at long range.

The way to differentiate between these possibilities might be to find out whether dark energy is changing over time. If it is, that would exclude the cosmological constant: as an inherent property of space, its density should remain unchanged. In most models of quintessence, by contrast, the energy becomes slowly diluted as space stretches – although in some it actually intensifies, pumped up by the universe's expansion. In most modified theories of gravity, dark energy's density is also variable. It can even go up for a while and then down, or vice versa.

The Dark Energy Survey, an international project in operation since 2013, is one attempt to fill in the gaps. It uses the 4-metre-wide Víctor M. Blanco telescope at the Cerro Tololo Inter-American Observatory in Chile, attached to a specially designed infrared-sensitive camera, to look for several telltale signs of dark energy over a wide swathe of the sky. For example, it catches many more supernova than before. The apparent brightness of each of these stellar explosions tells us how long ago it happened. During the time the light has taken to reach us, its wavelength has been stretched, or redshifted, by the expansion of space. Put these two things together and we can plot expansion over time.

The survey is also working on an intricate sky map that marks the positions of a few hundred million galaxies and their distances from us. Sound waves reverberating around the infant cosmos gave vast superclusters of galaxies a characteristic scale. By measuring the apparent size of superclusters, we can get a new perspective on the expansion history of the universe.

The map will also reveal dark influences on smaller scales. Dark energy hinders galaxies from coming together to form clusters. The survey team will count clusters directly and also follow their growth using an effect known as gravitational lensing, which happens when clusters bend light passing through them from even more distant cosmic objects. Early in 2021, the survey published the results of its first three years of data-taking. They were broadly in line with the standard model, while not yet being able to distinguish between different models of dark energy.

A full posse of dark energy hunters is due within the next few years in the form of huge new telescopes, such as the Thirty Meter Telescope in Hawaii and the Vera C. Rubin Observatory, the Extremely Large Telescope and the Giant Magellan Telescope, all in Chile. In 2023, the ESA and NASA plan to launch a dark-energy-hunting space mission called Euclid that will trace gravitational lensing and galaxy clumping to even earlier cosmic times.

This chase through space will be thrilling, but the quarry may still elude us. Say we find that dark energy maintains a near constant density over time. That would seem to support the cosmological constant, but it wouldn't rule out some quintessence fields that just happen to have a nearly constant density. Even if we find the dark energy density to be increasing or decreasing, we might not be able to tell whether that is due to quintessence or to some kind of varying gravity. Few imagine that this hunt will be over soon. ▮

THE HUBBLE TENSION

The standard cosmological model accounts for pretty much everything we observe in the universe at its grandest scales. But just recently, a problem has emerged – the model isn't accurately predicting the expansion of the universe.

THE success of the lambda-CDM model in reproducing the features of the cosmic microwave background (CMB) suggest a definitive test of consistency. You could take precise measurements of the universe's expansion rate when the radiation was released and use the model to wind forward and predict the current rate of expansion, known as the Hubble constant. "It's the ultimate end-to-end test of the universe," says Adam Riess, an astrophysicist at Johns Hopkins University in Maryland who led one of the teams that uncovered dark energy. "To go from the beginning to the end and have the two ends of the bridge that you are building meet up."

The trouble is that they don't meet. When we extrapolate forwards from the big bang using lambda-CDM, we get a lower rate of expansion than we do through astrophysical measurements of the distance to exploding stars in relatively nearby galaxies. The expansion of the universe is measured as the speed at which every million parsecs (Mpc) of space expands, a parsec being 3.26 light years. Working forward using lambda-CDM, cosmologists predict a Hubble constant of 68 kilometres per second for every million parsecs (km/s/Mpc). But looking at the rate of expansion today by measuring distances in space, astrophysicists get 73 to 74 km/s/Mpc.

This discrepancy is referred to as the Hubble tension. If lambda-CDM correctly describes the universe, it shouldn't be there. Most cosmologists, unwilling to give up on such an otherwise successful model, had assumed the tension isn't real – that the observations were wrong. But in 2019, a measurement made using a third method matched the higher, astrophysics-based value. A year later, the positions became even more entrenched when a new look at the CMB using the Atacama Cosmology Telescope in Chile bolstered the lambda-CDM prediction. "It is starting to get really serious," says Edvard Mörtsell, a cosmologist at Stockholm University in Sweden. "People must have really screwed up for this not to be real in some sense."

In the grand tradition of dark energy and dark matter, many solutions involve adding more unseen ingredients to lambda-CDM in the hope that this will increase the predicted expansion rate. But as well as fitting the Hubble constant, any model must correctly describe other observations, such as the rate at which galaxies form, the amount of galaxy clustering on various cosmological scales and the appearance of subtle ripples in the clustering of galaxies, known as baryon acoustic oscillations. Any changes that increase the Hubble constant quickly put these other predictions out of whack.

Another option is to tweak the behaviour of an existing component, for example by making the repulsive force supplied by dark energy stronger in the early universe. This eases the Hubble tension a bit, but nowhere near enough. The same goes for tweaking dark matter and for loosening assumptions about the even distribution of it throughout the universe.

A far more radical possibility is that something is out of whack in general relativity. One exotic proposal is that the problem lies in the way that the theory fails to account for the effects on gravity of the quantum mechanical property of spin in the matter that makes up celestial objects. Anything that affects the geometry of space-time will affect the expansion of the universe.

Or might the explanation still be entirely mundane? That is the suggestion of Wendy Freedman and her team at the University of Chicago in Illinois. To make local measurements of the Hubble constant, we use objects with known brightnesses to measure distances. Often, these are variable stars known as Cepheids. Freedman and her colleagues used a different type of star, called the tip of the red giant branch (TRGB) because of its place on charts of stellar evolution, to replace Cepheids – and the results determined from these stars matched the CMB measurements.

"It allows for the possibility that there is no tension, and it's just a matter of imperfect measurements," says Dan Scolnic at Duke University in North Carolina. We are still a way away from saying that, but the results provide a ray of hope that the tension might be resolvable after all, says Freedman. ■

BEYOND THE COSMOLOGICAL PRINCIPLE

An elegant assumption underpins our cosmic model: that everything looks the same everywhere. One controversial suggestion is that by abandoning it, we can banish dark energy from the cosmos – but that would involve jettisoning some other cherished assumptions, too.

OF THE simplifying assumptions used to make the fiendishly complex equations of general relativity tractable, one has a particularly distinguished pedigree.

When Nicolaus Copernicus laid out the Copernican principle in the 16th century, it was to say that Earth isn't the centre of the universe. In modern cosmology, it has morphed into the cosmological principle: that Earth is nowhere special at all. We see the universe from a representative standpoint, and draw conclusions that can apply everywhere else, too. The universe is homogeneous, looking roughly the same in all locations, and it is isotropic, looking roughly the same in all directions from any standpoint.

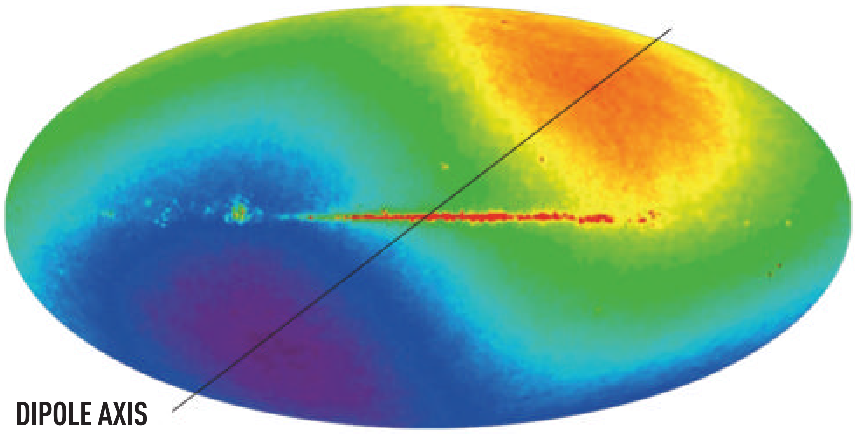
For some, these are simplifications too far. In the universe today, galaxies exist in clusters and filaments of matter distributed around the boundaries of huge, bubble-shaped voids. These voids have roughly one-tenth of the clusters' matter density, but account for more than 60 per cent of the universe's volume. "Everyone knows that the universe is inhomogeneous," says Thomas Buchert at the University of Lyon in France.

The mismatch is generally brushed aside using the concept of statistical homogeneity: the sort of universe we are looking for exists if we zoom out far enough. On scales of about 400 million light years, bigger than all the structures we see, voids and galaxy clusters average into uniformity. But we don't have a bird's-eye view of the universe on such scales. David Wiltshire, a cosmologist at the University of Canterbury in New Zealand, thinks that what is going on right in front of our eyes might be distorting our view.

He bases that on a well-known feature of the cosmic microwave background called the "dipole anisotropy" – that it actually appears overwhelmingly hotter in one direction than the other. This is generally explained by Earth's movement through space. Thanks to the Doppler effect, anything with a relative motion towards us looks hotter than it actually is and anything moving away looks cooler. Earth orbits the sun, the sun orbits the centre of the Milky Way, the Milky Way moves through our Local Group of galaxies and the Local Group is hurtling towards a massive concentration of more distant galaxies. Take account of the Doppler shifts created by all these motions and the hot and cold patches melt away.

Wiltshire takes issue with the last of the motions used to make the dipole anisotropy disappear: a movement at a speed of 635 kilometres per second of the entire Local Group towards a "great attractor" somewhere in the distant Hydra-Centaurus supercluster of galaxies. He and his colleagues think that the galaxies' movements make most sense if the Local Group isn't moving at all. Instead, the greater density of matter towards Hydra-Centaurus is slowing the universe's expansion along our line of sight, giving us the impression of such a movement.

A comparative void in the other direction, When astronomers look at the raw data of the cosmic microwave background, they see temperature variations in different parts of the sky. The principal variation is known as the dipole and is thought to be caused by the Milky Way's motion – the radiation is hotter than average in the direction of motion and colder in the opposite direction



meanwhile, is producing the opposite effect, causing an area of faster expansion behind us. The effects of the inhomogeneities along this axis are comparatively local, occurring on scales up to about 300 million light years, and only alter the universe's expansion rate by some 0.5 per cent. But they are sufficient to account for nearly all of the dipole anisotropy – and so colour our view of the entire universe.

Wiltshire emphasises that this is all just Einstein's relativity. "It is not controversial, it is not surprising, it is the norm, but it is just not the standard model," he says.

Physicist Paul Halpern at the University of the Sciences in Philadelphia, Pennsylvania, thinks such studies are worth pursuing, but warns of far-reaching consequences. "Until recently, no one doubted that the universe was homogeneous, so everyone just used the simplest models," he says. "Once you say that the universe can be very different in other parts of space, then you open up a can of worms. It would just be incredibly more complex to do cosmology."

The big question is whether inhomogeneity might kill off dark energy. The conclusion that the universe's expansion has begun to accelerate is intimately tied up with the assumption of uniformity. Buchert has developed models of an inhomogeneous universe that start from the self-same equations of Einstein's that everyone agrees on, but assume that the universe is fundamentally divided into voids and clusters of matter. When he does so, the universe no longer expands equally in all directions. Just as in Wiltshire's picture, a near void on one side of the sky makes the fabric of space in that region expand faster, whereas the gravity of a rich cluster of galaxies will make that area expand more slowly.

Crucially, though, as the universe has aged and gravity has clumped matter into ever larger galaxies and galaxy clusters, the voids between clusters have grown, and the universe has begun to unfurl ever faster in those regions. The result is an accelerating effect rather like that credited to dark energy, but without a hint of the stuff.

Natural it might be, but an inhomogeneous, lumpy universe would work very differently to the one we think we know. For a start, Einstein showed that space and time are conjoined entities, so if you allow space to expand at different rates in different places, you must accept that clocks will tick at different speeds, too. That means even such a fundamental property as the age of the universe wouldn't be constant all across the cosmos. Measure it from within a dense cluster and you will get one answer; measure it from within a void and you will get another.

Wiltshire countenanced this possibility in earlier work on a theory he calls the "timescape". This suggests that the age of the universe could be as much as 18.6 billion years in places where a low density of matter means the clock has ticked particularly fast. Our own smaller estimate of the universe's age is a natural consequence of sitting in an area of unusually high density: a galaxy.

Most cosmologists remain unconvinced. Alexander Kashlinsky at NASA's Goddard Space Flight Center in Greenbelt, Maryland, argues that the peerless agreement between the standard-model predictions and observations of how galaxies form, for instance, means it must be something close to the truth. And the factor of 10 difference in matter density between galaxy clusters and voids isn't enough to justify jettisoning the standard model. "The inhomogeneities are at such a low level that the overall description of the homogeneous universe that we use is a very good assumption," he says. He would rather keep things like dark energy – even if to explain them we must look to theories that leap beyond Einstein. ▮

CONTEMPLATING THE FATE OF THE UNIVERSE

General relativity implies that some day in the distant future, the universe will end. Katie Mack has made it her business to understand how.

PROFILE
KATIE MACK

Katie Mack is a cosmologist at North Carolina State University and author of *The End of Everything (Astrophysically Speaking)*. Her Twitter account @AstroKatie has over 400,000 followers

Do we know how the universe will end?

The [scenario] that I think is most likely based on current data is called the heat death. If the universe is expanding, and if its expansion continues to speed up, then space will get more and more dilute over time, which is to say there will be more and more space between each galaxy. Eventually, space gets so dilute that matter in the universe becomes less and less important. Galaxies stop colliding with each other, so they aren't bringing in enough gas to make new stars and the old stars are burning out. Even black holes will disappear.

As time goes on and things decay, that increases entropy, which is the disorder of the universe. If you leave the universe alone for long enough and it's decaying over time, you end up in this maximum entropy state where all that's left is this tiny amount of background radiation known as waste heat. Once you get to maximum entropy, nothing else of importance can really occur.

Of all the possible scenarios, which is your favourite?

My favourite scenario is vacuum decay. It's this idea that's been around since the 1970s that our universe might not be entirely stable. It's all based on the Higgs field, which is a field related to the Higgs boson, the particle that was discovered at the Large Hadron Collider (LHC) at CERN [the particle physics laboratory near Geneva, Switzerland] in 2012. The energy of the Higgs field determines whether the universe is in its lowest possible energy state, known as a true vacuum, or a false vacuum, which is a slightly higher energy state.

The conditions in the early universe determined which state the Higgs field would be in, and if it's in

NERISSA ESCANLAR/EARTH-LIFE SCIENCE INSTITUTE

BIG CRUNCH, BIG FREEZE OR BIG RIP?

Before the discovery of dark energy, the future of the universe depended on geometry. Either the cosmos was closed and would collapse in on itself in a "big crunch" or it was open and would expand forever. Now, however, cosmology's standard model assumes that we live in a flat universe that, thanks to dark energy, will expand eternally.

If dark energy is nothing more exotic than a cosmological constant, meaning it doesn't fluctuate over time, then the heat death of the universe (or the "big freeze") will be the outcome. But it isn't the only possibility. An alternative is the "big rip". Here, the dark energy keeps getting stronger and the expansion of the universe keeps accelerating. This is, in some ways, the most exciting option on the table. Even gravitationally bound objects like galaxies can eventually get pulled apart, as dark energy overpowers the gravity holding them together.

a false vacuum state, then it could spontaneously transition to a true vacuum. That would rewrite all the laws of physics and constants of nature as we know them.

Physicists call this process vacuum decay. In the universe after vacuum decay, the new laws of physics would make it impossible for, say, molecules to exist, because the way that atoms interact with one another would be different. Space itself would be unstable, and eventually everything would collapse into a black hole.

What could make this happen?

Imagine balancing a glass right on the edge of the table. It's fine right now, but it would prefer to be on the floor because that's the lowest energy state, and something could happen at any time that could push it over. Similarly, it's possible that our universe prefers a different value of the Higgs field and the slightest touch could knock it over. Like the glass, it would be more stable, but it would be broken.

There are two ways for this to happen. One would be that something disturbs the Higgs field. That would have to be an extremely high-energy event, much higher energy than we can even imagine. When the LHC first started up, there was some worry that its collisions could create a high enough energy to disturb the Higgs field, but they are nowhere near powerful enough to do that.

What is the other way?

The other idea is that the transition could happen spontaneously through a phenomenon known as quantum tunnelling. If you have a particle on one side of a wall, quantum mechanics says it's possible › for the particle to spontaneously appear on the other side. In theory, if you put a glass on the edge of the table, all its constituent particles could align and allow it to just spontaneously quantum tunnel to the floor. It's extremely unlikely to happen, but we can't rule it out.

If something like this happened to our universe, a bubble of the new vacuum would spontaneously form within it: a region where we can't exist, because our molecules would fall apart, and space itself collapses. And it would expand at roughly the speed of light. It would plough through the universe and destroy everything within it. If it got you, you wouldn't see anything or feel anything: you're just done. It's this very dramatic way to destroy the universe.

Should we be worried about vacuum decay?

There are several reasons not to worry – for one, the false vacuum is predicted to stay stable for way longer than the current age of the universe – but physicists are paying a lot of attention to it now because our experiments do suggest that it's possible.

What comes after the end of the universe? Is it just nothingness?

I define the end of the universe as the end of our observable universe – the volume of space that we can interact with, that has any impact on us or that we have any impact on. If everything in that region is destroyed, I rate that as the end of the universe. It doesn't mean that there couldn't be more space beyond that where more things continue, or another universe after ours, but for us, the end of the observable universe is the end.

How do you feel about the end of the universe?

When I was putting together my book, I interviewed a bunch of other cosmologists and astronomers about how they feel about the end of the universe: are they sad about it, or have they come to terms with the fact that we won't go on forever?

A few people said that it was really sad. One person said that when she gives lectures about the heat death, people sometimes cry.

I haven't really decided how I feel about it yet. I'm still kind of trying to wrap my head around it in some meaningful way. I am somebody who is not at all comfortable with the idea that I will die some day, for example. Intellectually, I know that that's true, but it's also terrifying. So the idea that the whole universe will die some day, that everything I love and care about will be over, is hard to wrap my head around.

Does thinking about things on this massive scale help you put daily troubles in perspective?

There's something about studying the forces of nature that changes how you view everyday life. It doesn't so much make everything insignificant, but it makes clear how little control we really have.

We live in a society with the illusion of control, and there's a sense of security in how much we've altered our surroundings and built a world that suits us. But when you get to the bigger picture, we're this tiny little speck of dust adrift in the cosmos with no say over what happens to our cosmic environment or the universe as a whole, however much we eventually come to understand it.

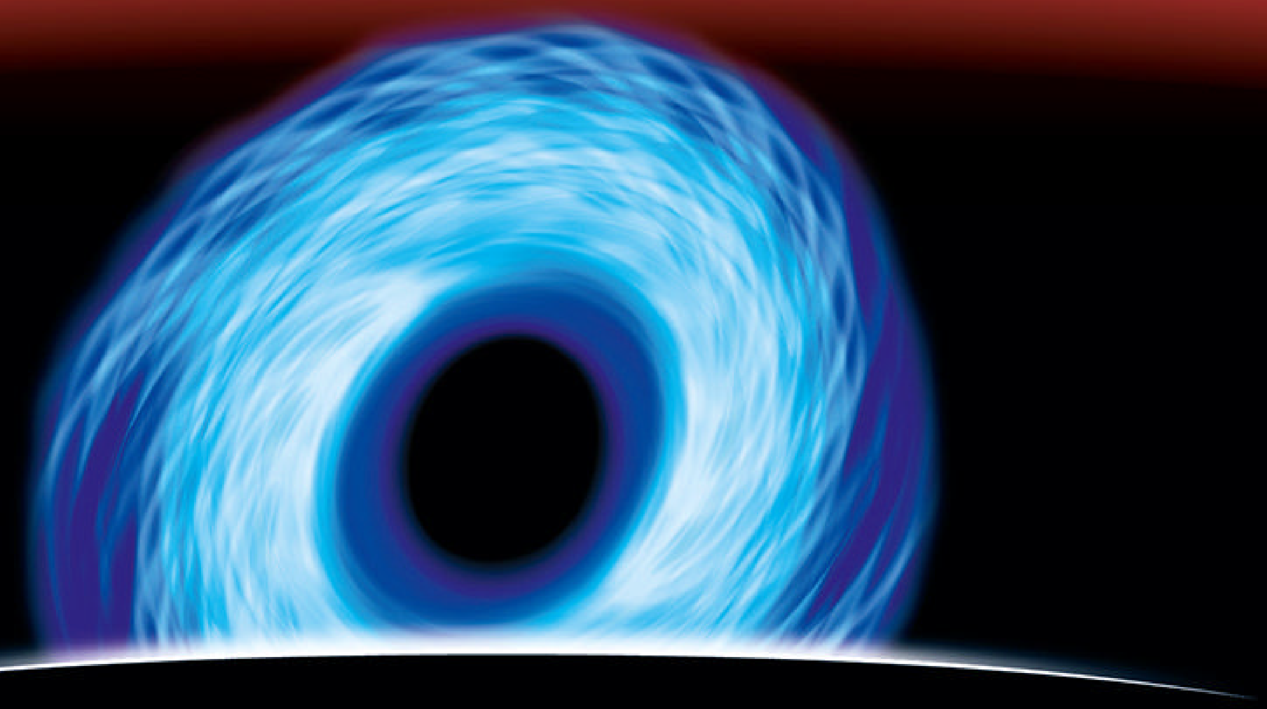
Studying these kinds of things, it's not like it's reassuring at all, but it chips away at the illusion of control in a way that lets you step back a little bit. Sometimes things are just going to happen and the universe doesn't care about any of it.

All we can do is make the best out of what we have. There's some amount of comfort in the fact that we're all in this together, at the mercy of some of these bigger forces, and that's OK. ■

New
Scientist
Academy

MASTER
THE
UNIVERSE

Fill the black holes in your understanding of the cosmos with an out-of-this world learning experience...

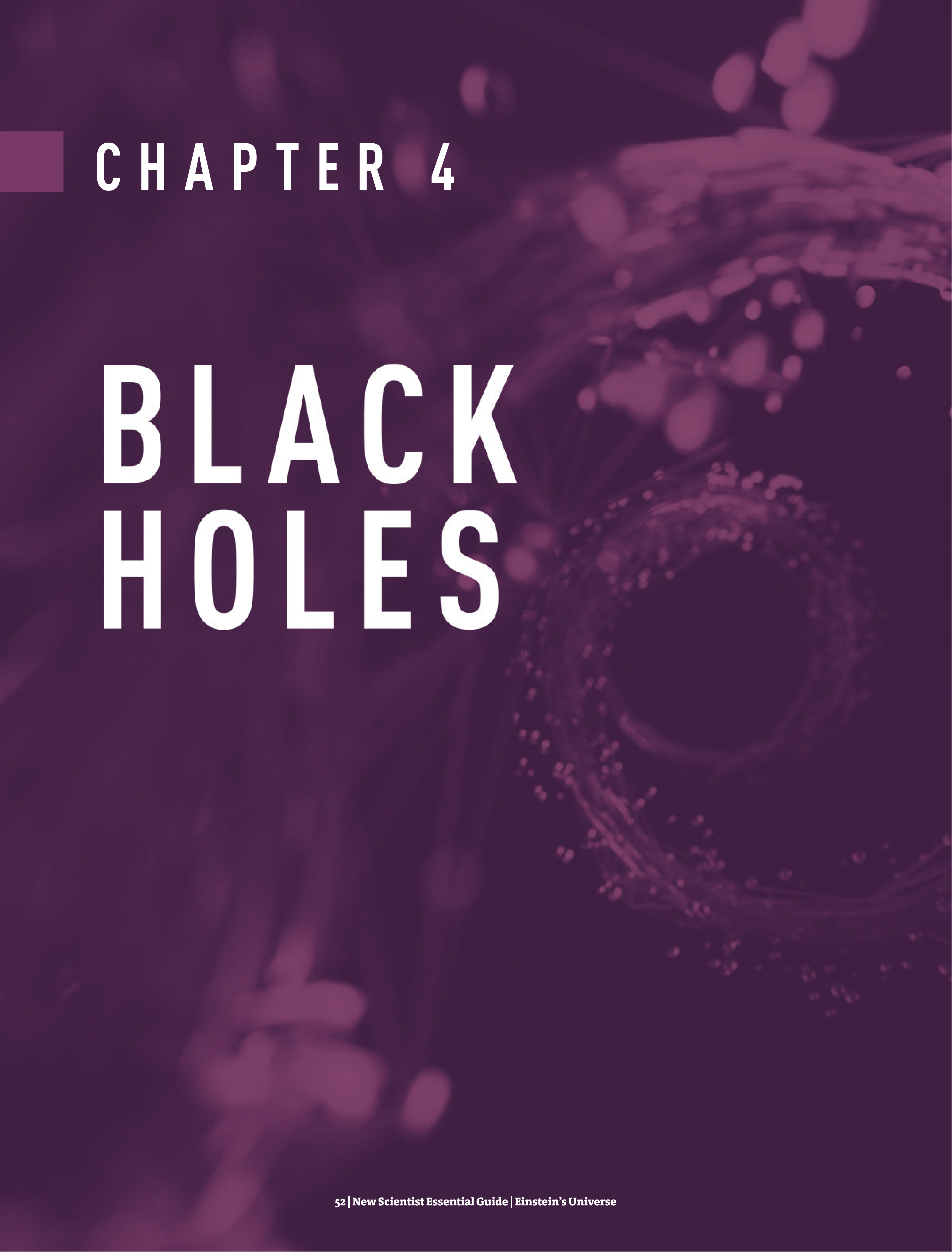


Learn from the experts at your own pace with New Scientist Academy's immersive interactive courses. Discover the Beyond Einstein course at newscientist.com/beyond-einstein

New
Scientist
Academy

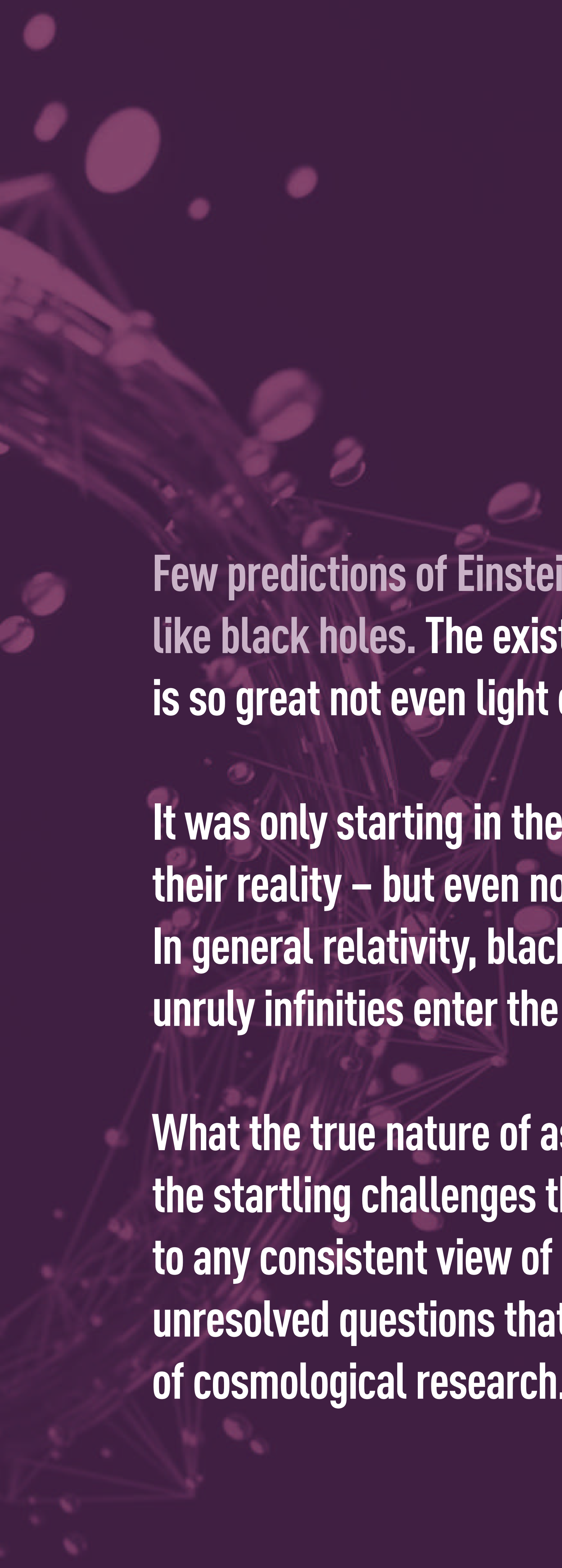
Master the universe.
Make sense of general relativity

Beyond Einstein



CHAPTER 4

BLACK HOLES



Few predictions of Einstein's relativity capture the imagination like black holes. The existence of these monsters, whose gravity is so great not even light can escape them, was long disputed.

It was only starting in the 1960s that we became convinced of their reality – but even now, we aren't sure what we have seen. In general relativity, black holes represent "singularities" where unruly infinities enter the mathematics, meaning all bets are off.

What the true nature of astronomical black holes is, and the startling challenges that the theoretical objects pose to any consistent view of how the cosmos works, are unresolved questions that take us to the cutting edge of cosmological research.

THE REALITY OF BLACK HOLES

Over the past few decades, a battery of evidence has left us in little doubt that black holes exist – whatever they are.

PREVIOUS PAGE: JUST_SUPER/ISTOCK
PHILIPP TUR/ISTOCK

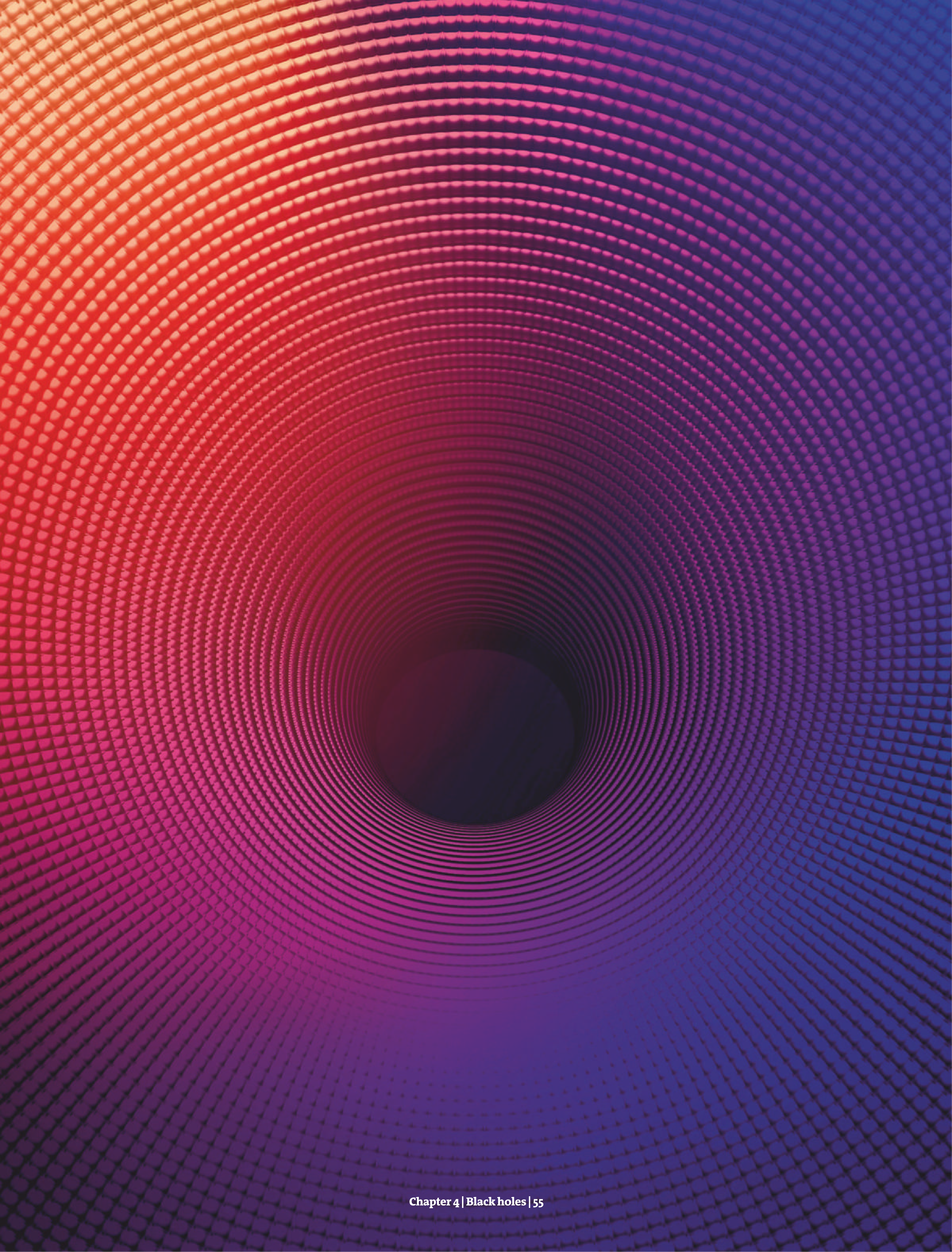
IT WAS while serving in the German army on the Russian front in the winter of 1915 to 1916 that physicist Karl Schwarzschild sent Albert Einstein some papers. He had solved Einstein's equations of general relativity for the first time and shown what happens to space-time inside and outside a massive object – in this case, a perfectly spherical, non-spinning star. Einstein was thrilled.

He wasn't so thrilled with a prediction that eventually emerged from Schwarzschild's work. Make a star compact enough and it could develop a gravitational pull so great, and warp space-time so much, that even light wouldn't get away.

In fact, he wasn't the first person to toy with the idea of an object so massive that not even light can escape the pull of its gravity – that was English geologist John Michell back in 1783. General relativity gave the idea new impetus. Within a year of his exchange with Einstein, Schwarzchild was dead, so it was left to others to work through the details of the curious compact objects he had envisaged, the surfaces of which became known as "Schwarzchild singularities".

Chief among them was a young Indian physicist named Subrahmanyan Chandrasekhar. Whiling away an 18-day voyage aboard a steamer to take up a scholarship at the University of Cambridge, he worked on the properties of highly compact white dwarf stars. He found that if they had more than 1.4 times the sun's mass, they would implode under their own gravity, becoming so dense that they would form Schwarzschild singularities.

This didn't go down well. At a meeting of the Royal Astronomical Society in 1935, Arthur Eddington – the ›



eminent astrophysicist who had verified general relativity's predictions during a solar eclipse in 1919 – declared that "there should be a law of nature to prevent a star from behaving in this absurd way". In 1939, Einstein himself published a paper to explain why Schwarzschild singularities couldn't exist outside the minds of theorists.

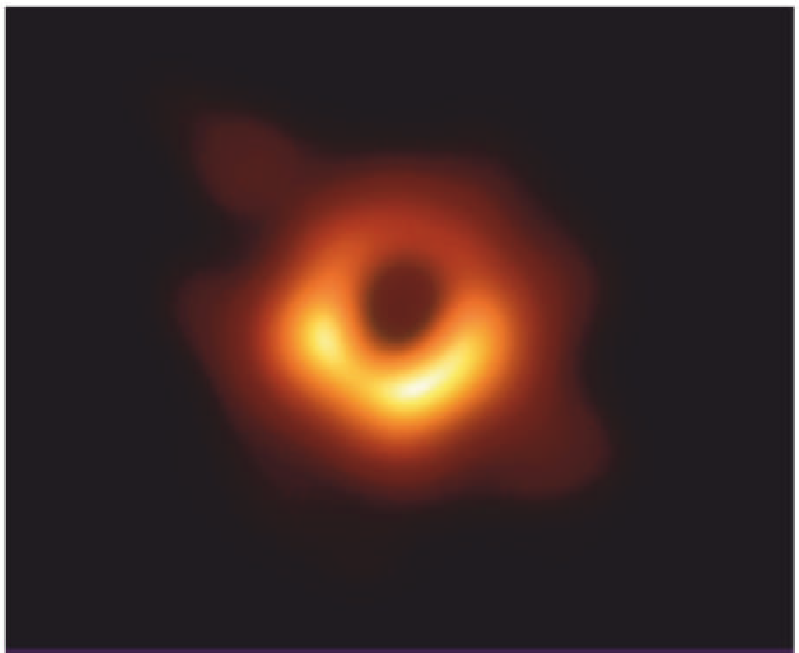
The impasse remained until the 1960s, when physicists such as Roger Penrose proved that black holes – a term coined at about this time, probably by astrophysicist John Wheeler – were a seemingly inevitable consequence of the collapse of massive stars. At a black hole, physical quantities such as the curvature of space-time would become infinite, and the equations of general relativity would break down.

Not only that, but a black hole's interior would be permanently hidden behind its "event horizon", the surface of no return for light. Nothing happening in the interior could influence events outside, because no matter or energy could escape. "The first major paradigm shift was the understanding that these solutions [of general relativity] are meaningful, and that there is a notion called a horizon, and that it is a causal barrier separating the inside from the outside," says theorist Don Marolf at the University of California, Santa Barbara.

The process of collapse would also destroy every characteristic of the original star except its mass, spin and electric charge: everything else is radiated away as gravitational waves. The resulting black hole is said to "have no hair" – to bear no trace of its former existence.

The notion of such stellar-mass black holes seemed no more than a mathematical trick until 1968, with the discovery of pulsars. Pulsars are rapidly spinning neutron stars that contain about as much matter as our sun in a volume about that of a large mountain on Earth. That puts them very close to the critical density at which gravity would overwhelm them and they would collapse into black holes. A neutron star could gain enough extra mass to do the trick, either by accumulating matter from interstellar space or by its gravitational pull stripping gas from a companion star.

This made the possibility of astronomical black holes respectable, and although they give out no light, there are several ways astronomers can search for them.



THE FIRST BLACK HOLE IMAGE

In April 2019, the Event Horizon Telescope collaboration unveiled the first direct image of a black hole (see above), the product of a network of telescopes around the globe wired up to turn Earth into an enormous radio dish.

This image shows the large black hole in the centre of another galaxy called Messier 87, which is 55 million light years away. The black hole is about 7 billion times the mass of the sun and some 100 billion kilometres wide – about 22 times the average distance between Neptune and the sun.

Nothing we can see in the image is coming out of the black hole. The black hole's event horizon is in the black area – the shadow of the black hole against the bright, luminous material circling it and eventually falling in. The asymmetry in the image is caused by the black hole's rotation – the light that is coming towards us appears brighter, and the light moving away doesn't seem as bright.

EHT COLLABORATION

Their existence can be inferred by their powerful gravitational pull on nearby stars. In some cases, stars are found to be orbiting an invisible partner, and if calculations show that the partner has more than a certain mass, it is probably a black hole.

A black hole's intense gravity also tends to attract gas and dust. Friction heats up this material, causing it to release vast amounts of radiation, which telescopes can detect. Light from stars that lie behind a black hole as seen from Earth should be deflected by its gravity. This process is called gravitational lensing, and the measurements of the deflection of light can again be used to infer the existence of the black hole.

In 1970, astronomers observing a compact object in the constellation Cygnus saw jets of X-rays consistent with theoretical predictions of radiation streaming from hot matter spiralling towards an event horizon. Black holes were soon invoked to explain another puzzling discovery of the 1960s: quasars. These are the energetic cores of some galaxies, which produce enormous amounts of energy from a region of space no bigger across than our solar system.

The kind of black hole involved would contain maybe a few hundred million times our sun's mass, far more than a stellar-mass black hole. Such supermassive black holes could arise simply because too many stars got too close together in the core of a galaxy. We now think most galaxies have one at their heart.

That includes our Milky Way, the existence of whose central black hole, Sagittarius A*, was proved by Reinhard Genzel and Andrea Ghez by following the motions of nearby stars over decades – a feat for which they received a share of the 2020 Nobel prize in physics.

↓
There's an interview with Andrea Ghez on page 64 later in this chapter

More recently, observations of gravitational waves – ripples in space-time given out when two black holes collide and merge – have provided more evidence of black holes' existence. Finally, in 2019, a black hole was directly pictured for the first time (see "The first black hole image", left). That they exist is now hardly in doubt – and that's where the true problems start. ▮

ESSAY

THE GREAT BLACK HOLE PARADOX

Black holes devour everything that comes too close. Explaining what happens to that stuff has exposed a fundamental gap in our theories of the cosmos, says Paul Davies.

A BLACK hole's defining feature is its event horizon, the boundary inside which gravity is so strong that light can't escape. As nothing can go faster than light, this means anything crossing the event horizon is irreversibly lost to the outside universe.

At least, that is the case on a simple reading of general relativity. This also says that hidden from view in the heart of a black hole is a "singularity", an infinitely warped edge or boundary of space-time where the laws of physics break down. Any matter that hits a singularity – and, crucially, any information encoded in that matter, for example how the molecules in a cloud of gas are distributed – must disappear from space-time.

This is a challenge to traditional views in physics about time and irreversibility. Compare the fate of an encyclopedia thrown into a black hole with one put into an incinerator. With the incinerator, if you knew the precise state of every molecule and every photon radiated as heat, you could, in principle, "run the

Stephen Hawking's work hugely increased the mystery of black holes

KEVIN DIETSCH/UPI/ALAMY STOCK PHOTO

movie backwards" and recover the information contained in the encyclopedia. Not so with a black hole. The information loss appears to be absolute and objective: there is no rewind button.

The puzzlement really turned up a notch, however, in early 1974, when Stephen Hawking delivered a famous lecture at what is now the Rutherford Appleton Laboratory near Oxford, UK. I was there. Hawking announced that black holes aren't totally black, but glow faintly because of the effects of quantum particles that pop up out of the vacuum near its event horizon and are radiated away. The process of emitting "Hawking radiation" slowly sucks energy from the black hole, so it gradually shrinks over an immense length of time.

This was a sensational claim. The Hawking effect was puzzling on several levels, but one question stood out: if a black hole goes on shrinking, does it eventually totally disappear – and if so, what happens to all the stuff that fell into it?

Hawking derived his result by appealing to quantum mechanics. Its laws are time-symmetric, so in theory you should be able to gather all the information encoded in the Hawking radiation and work backwards to the starting state, just as with an incinerated encyclopedia. But Hawking's calculations showed that the radiation produced by a black hole is precisely "thermal" – entirely random – containing no information whatsoever about what fell into the hole originally.

This is the basis of what is known as the black hole information paradox. The laws of quantum mechanics say that information can't be destroyed. General relativity, by introducing black holes, apparently says it must be.

I first discussed this clash with Hawking in a hotel room in Boston in the 1970s, where we had both travelled for a conference. At that time, Hawking, who was steeped in the general theory of relativity and its predictions about black hole singularities, thought the paradox indicated that quantum mechanics must break down in black holes. He published a paper claiming as much, containing the memorable aphorism – echoing Einstein's criticism of quantum

PROFILE
PAUL DAVIES

Paul Davies is a cosmologist and physicist at Arizona State University in Tempe. His research interests span quantum gravity and black holes, the nature of time, the origins of life and the evolution of cancer. He is the author of more than 30 books

theory that "God does not play dice" – that "not only does God play dice, but… he sometimes throws the dice where they cannot be seen".

Over the subsequent decades, however, many physicists have come to believe that quantum mechanics is sacrosanct, and that the lost information must somehow be returned to the outside universe. That is especially true among string theorists, whose efforts to construct a quantum theory of gravity are rooted in the standard rules of quantum mechanics. After wobbling for years, Hawking finally concurred. What went into the hole, he declared, must come out – in one form or another. But how?

In the absence of a satisfactory theory of quantum gravity, Hawking's original calculation was, crucially, "semiclassical". It applied quantum mechanics to fields such as electromagnetism around the black hole, but not to the black hole's own gravitational field. There is general agreement that such an approximation will break down, and quantum gravity effects must kick in, at the Planck scale of about 10^{-33} centimetres. This is a number calculated by combining Planck's constant, which sets the strength of quantum effects, and Newton's gravitational constant, which determines the strength of gravitation. The hope was that when a black hole shrank to such a size, new effects would emerge to solve the paradox.

→
Chapter 6 has more on quantum gravity

But as Don Page, a former postdoctoral collaborator of Hawking's, pointed out in a major twist back in 1992, we can't sweep the problem under a Planck-scale carpet. That is because of entanglement, the quantum phenomenon described by Einstein as "spooky action at a distance". It says that if a pair of particles – for example photons of light – is created from the quantum vacuum, and the particles fly off in opposite directions, they remain intimately linked in their properties. Independent measurements performed simultaneously on the two particles will uncover that link.

Entanglement is a much-studied quantum phenomenon, because it forms the basis for the design of quantum computers. Applied to Hawking radiation, pairs of entangled particles are created near a black hole, with one escaping and the other falling down the hole. Their entanglement implies a subtle residual connection reaching across the event horizon.

In thermodynamics, physicists quantify lost or hidden information in terms of entropy, a general measure of disorder. When information goes down, entropy goes up, and vice versa. Every time a pair of photons is produced and one slips over the event horizon, "entanglement entropy" increases. When the Hawking effect starts out, the entanglement entropy is zero, but it rises steadily as more and more particles get created and separated by the horizon.

Page realised that this inexorable rise must have a limit. As originally suggested by Jacob Bekenstein in 1972, and confirmed by Hawking a couple of years later, a black hole possesses a total entropy proportional to its surface area. As a black hole evaporates, its surface area shrinks, and so does its total entropy. Thus, the entanglement entropy rises and the total entropy falls until, about halfway through the evaporation process, they become equal.

At that point, something changes. Entanglement entropy can no longer go up, but falls with the total entropy as the hole continues to shrink. This loss of entanglement entropy implies the appearance of information. But where? As departures from randomness in the Hawking radiation; that is to say, correlations between particles within it. These correlations grow over time as the black hole shrinks towards its eventual demise. According to Page's analysis, the original entanglement between pairs of outgoing and ingoing particles reappears as entanglement between outgoing particles – specifically, between particles emitted at earlier times and those emitted at later times. Entanglement in space becomes entanglement in time.

Significantly, the turnover point occurs when the black hole is still a macroscopic, possibly huge object, very far from the Planck size at which quantum gravity can't be ignored. The build-up of correlations in the outgoing Hawking flux would seem to be a neat way › out of the information paradox. Information in does indeed equal information out, but it is concealed by being spread over time. If this is correct, the reversibility of the laws of physics is preserved by the black hole evaporation process.

That is all well and good, but to buy this argument, you must conclude that there is something missing from Hawking's original calculations, which say there is no entanglement or information in the radiation from the black hole. And there is no agreement on where the flaw might lie. Attempts to provide an answer so far have either appealed to idealised special cases or descended into speculative mathematical backwaters with only a tenuous link to reality. They provide at best circumstantial (and entirely theoretical) evidence that the information about the material that went into the black hole reappears in some guise in the Hawking radiation.

One such idea is that the entanglement between pairs of particles produced near a black hole's event horizon somehow gets erased before one falls down the hole. This entanglement destruction would release a vast amount of energy, resulting in an intensely destructive, incinerating surface known as a firewall encircling the event horizon. This firewall should produce conspicuous effects outside the black hole, but it contradicts a fundamental tenet of general relativity that the event horizon has no special local properties: it just marks the boundary where the strength of the black hole's gravitational field becomes great enough that light can't escape. The firewall prediction also comes from the questionable practice of considering particles as little packets of localised energy. Direct calculations of the quantum energy density round a black hole, first done in the 1970s, show it to be smoothed out and continuous at the event horizon.

Some theoretical physicists retain the belief that only a fully worked-out theory of quantum gravity will produce a resolution to the paradox. Such a theory will probably include not just intense space-warping effects, but a feature known as topology change. Way back in the 1950s, John Wheeler pointed out that, on the Planck scale, quantum vacuum fluctuations would be so powerful that they would bend space-time into a sort of foamy structure – a frenetically shifting landscape of wormholes and bridges connecting different regions. Wheeler thought that, in place of a point-like singularity at a black hole's centre, there should be a foamy blob. Topology change might also create a type of tunnel or wormhole linking the interior of a black hole with another universe or a distant region of our own universe, an idea first suggested by Wheeler and championed by several others.

If that were the case, you could fall through a black hole and come out in a completely different space. Then there needn't be an information paradox. The information about the infalling matter could simply traverse the wormhole and continue to exist in the other region of space-time. As long as we humans are restricted to "our" space-time region, information is lost, but taking a God's-eye view, information would be conserved.

The possibility that wormholes might connect the interior of black holes with another area of our own space-time outside the hole, allowing information to sneakily leak back out, is the basis of renewed claims recently that the black hole information paradox is close to resolution. But these calculations, as is so often the case, rely on highly idealised analogues of real black holes and involve layers of simplifying assumptions, so it isn't clear how relevant they really are.

There is a more general concern, too, about the uncritical application of quantum mechanics to the black hole evaporation process. Calculations tend to assume that the black hole and its products form an isolated system, which is obviously unrealistic. Quite apart from the disturbing effects of the rest of the universe, there is a fundamental question concerning what we mean by information. For information to be extracted from a quantum system, a measurement has to be performed by an external system. The very act of measurement breaks the time symmetry of quantum mechanics in a process sometimes described as the collapse of the wave function. So if "information" is treated as something that could actually be gleaned from a measurement performed on Hawking radiation, the rewind button is destroyed as soon as that measurement is made.

The black hole information paradox is an inconvenient truth at the heart of physics, yet it has spurred a rich variety of theoretical investigations that have pushed the frontiers of the subject in important new directions. When Hawking announced his black hole evaporation result, it established a link between quantum mechanics, gravitation and thermodynamics. This is surely an important clue, and suggests that the resolution of the paradox – which is undoubtedly out there – lies in a revolution that unites our understanding of all three of those elements. Almost half a century on, however, we are still waiting for that revolution. It might take another Stephen Hawking to start it. ∎

↓
Turn to page 66 for physicist Carlo Rovelli's answer to the black hole paradox

WHAT HAPPENS IF YOU FALL INTO A BLACK HOLE?

No one knows for certain what goes on beyond a black hole's event horizon – but we can sketch out some of what would happen to you if you were to stray too close.

For a start, a black hole's gravitational pull is so massive that time itself would start to warp. You wouldn't feel anything different as you fell in, but to anyone watching, you would appear to slow down. "That's when the universe starts to go bizarre on you," says Priyamvada Natarajan at Yale University.

Circling down the drain of this cosmic plughole, all the photons being pulled alongside you would create a stream of blinding light orbiting a hole of total blackness. Two freaky effects would colour the final approach: a looming darkness would wash over your eyes as the black hole seemed to grow in size much more quickly than you would expect, and the surrounding stars would start to distort and bend.

"By the time you are able to see the event horizon," says Natarajan, "you can see starlight bend around it." Curved streaks of light would wrap around the black hole, but as you accelerated through the point of no return, the intensity of the gravitational field would change the nature of the light you could see. If your neck muscles had the strength to let you take one last look over your shoulder, all the starlight behind you would appear to come together to form a single reddish dot.

Meanwhile, the quiet of outer space would turn to total darkness and you would feel that you were falling downhill, says Natarajan – "except downhill is everywhere".

It wouldn't be a pleasant experience. The gravity inside a black hole increases so quickly that it wouldn't just crush you, but pull apart every bit of your body at different speeds, resulting in "spaghettification".

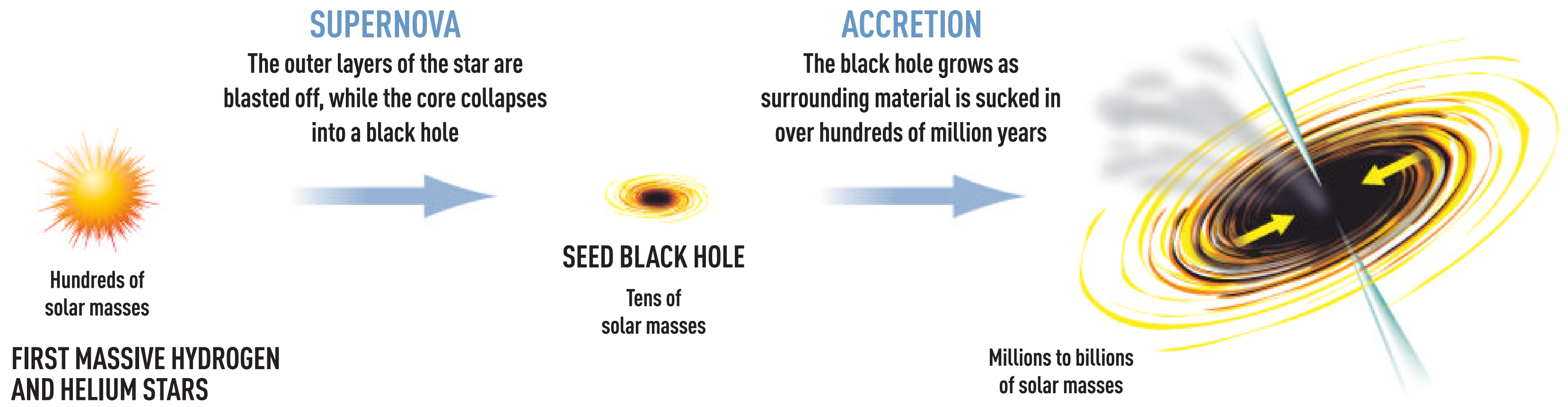
If you fell in feet first, your ankles would stretch away from your knees before your neck elongated into a strand of linguine, but the difference in time would be small enough that you probably wouldn't notice. "It would happen in the blink of an eye, which may not be the best expression since your eyes are going to pop out," says Natarajan.

After this, your fate is a matter of pure conjecture. Some physicists have suggested that you could cling on inside a black hole as it evaporates via Hawking radiation, before getting belched out in its dying breath.

Others say that by the time the black hole got small enough to vomit out its last meal, there wouldn't be anything left to expel. Or perhaps you might escape via a white hole, or a wormhole passage to another part of the universe. Perhaps it's better just not to test it.

VECTARAY/DRGROUNDS/ISTOCK

The old picture of supermassive black hole formation has the cosmic giants feeding constantly on surrounding gas over hundreds of millions of years


SUPERNOVA
The outer layers of the star are blasted off, while the core collapses into a black hole
ACCRETION
The black hole grows as surrounding material is sucked in over hundreds of million years
Hundreds of solar masses
FIRST MASSIVE HYDROGEN AND HELIUM STARS
SEED BLACK HOLE
Tens of solar masses
Millions to billions of solar masses

WHAT MADE BLACK HOLES SO BIG?

Supermassive black holes have all the space-warping strangeness of their smaller kin, but they hold another level of mystery. Nobody can explain the genesis of the giants – unless, perhaps, it points back to an origin right at the beginning.

WHATEVER made supermassive black holes, it made an army of them. Observations of stars whirled around by powerful gravity suggest that there is a huge black hole at the heart of almost all large galaxies, including our own. The Milky Way's is 5 million times the mass of the sun. The one imaged at the centre of the giant elliptical galaxy Messier 87 in 2019 has more than 6 million solar masses. Its event horizon is nearly five times as wide as the orbit of Neptune.

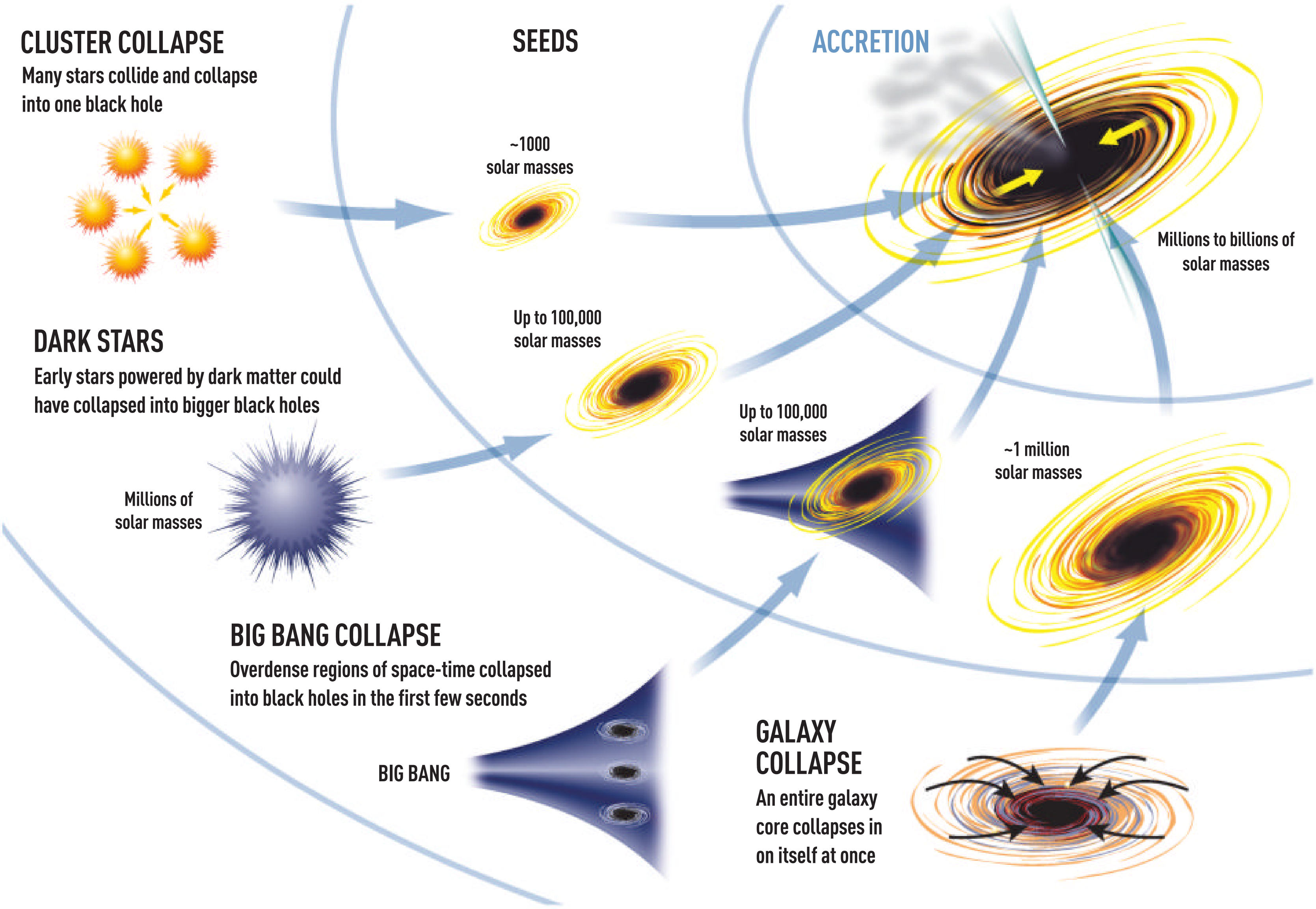
The conventional tale of supermassive black holes starts with a smaller seed maybe a few hundred times the mass of the sun. As this seed gorges on gas, it sinks towards the centre of its galaxy, eventually becoming the powerful heart of a quasar. But theory suggests a limit to this process. The more gas a black hole gulps, the more light and other radiation that shines out. Eventually, the flood of light grows so fierce that it sweeps away any incoming gas, cutting off the supply. Theory, backed by the behaviour of black holes nearby, says that a black hole can double in size once every 30 million years at most.

The problem lies in the fact that we have now seen several large supermassive black holes way out in the distant cosmos within the first billion years after the big bang – and those black holes simply shouldn't have been able to become so big in the time available to them.

One possibility is that supermassive black holes began not with single stars, but many. "We know that early in the history of the universe, stars tended to form in bursts – regions that were spectacularly active," says Fred Rasio at Northwestern University in Evanston, Illinois. Rasio's educated guess is that they might well collapse to form a black hole of perhaps a few thousand solar masses, a weightier starting seed. The catch is that we can't find similar middleweight black holes in star clusters today.

Another possibility is that a bigger seed might come not from a collapsing star, but from the collapsing

Quasar observations suggest supermassive black holes already existed in the early universe – meaning we need to explain how their seeds got big so quick


CLUSTER COLLAPSE
Many stars collide and collapse into one black hole
SEEDS
ACCRETION
~1000 solar masses
Millions to billions of solar masses
DARK STARS
Early stars powered by dark matter could have collapsed into bigger black holes
Up to 100,000 solar masses
Millions of solar masses
Up to 100,000 solar masses
~1 million solar masses
BIG BANG COLLAPSE
Overdense regions of space-time collapsed into black holes in the first few seconds
BIG BANG
GALAXY COLLAPSE
An entire galaxy core collapses in on itself at once

heart of a galaxy. This was originally suggested as an outside possibility in 1978 by Martin Rees at the University of Cambridge, but it turns out it isn't easy to cram so much matter into a galaxy's heart. A more exotic idea is that early "dark stars" powered by dark matter might have collapsed into bigger black holes.

←
Page 39 has more on dark matter

But perhaps the most radical suggestion is that giant black holes were forged directly in the fires of the big bang. In some models of the process of cosmic inflation, a period of faster-than-light expansion most cosmologists think happened in the universe's first instant, tiny fluctuations in the universe's density would have ballooned into small areas of extreme density that could have pulled in beams of ambient light before collapsing into black holes. Such "primordial" black holes might solve the problem of the over-supermassive black holes. "If you start out with a million-solar-mass black hole in the early universe, it's easy to get to a billion," says Bernard Carr at Queen Mary University of London.

Primordial black holes are popular among theorists for other reasons, too. For one thing, they might make up the universe's missing dark matter. The easiest way to catch them red-handed would probably be by spotting gravitational waves given out as they merge. We could alternatively look for radiation emitted by matter falling into them, or use gravitational lensing, a phenomenon whereby massive objects stretch and distort the light passing near them.

Seeing one for certain would mean peering back billions of years into the first few hundred million years of the universe, as systems with non-primordial black holes wouldn't have been possible then. Beyond finding a small, evaporating black hole or spotting one in the early universe, though, there are very few ways to prove that an observed black hole is primordial. Carr, who has devoted his career to these black holes, puts the probability that they are real at between 20 and 50 per cent – and so the mystery of how supermassive black holes got so vast so fast continues. ▮

HOW I PROVED SUPERMASSIVE BLACK HOLES ARE REAL

Twenty years ago, Andrea Ghez set out to show that there is a black hole at the centre of our galaxy by watching stars orbit it – and succeeded in what many thought an impossible task.

PROFILE
ANDREA GHEZ

Andrea Ghez is an astronomer at the University of California, Los Angeles. Together with her collaborator Reinhard Genzel, she won a share of the 2020 Nobel prize in physics for her studies of Sagittarius A*, the Milky Way's central black hole

Why did you start studying supermassive black holes?

I think it was the early moon landings that first got me interested in astrophysics and thinking about the scale of the universe. What was bothering me was boundaries – the beginning and end of time and the boundaries of space. Black holes really capture a lot of those problems with space and time, especially with how general relativity and quantum mechanics come together, so I think that's originally what got me interested in black holes. They really represent the boundary of our understanding of how the universe works.

These monsters live at the centres of galaxies. What is that environment like?

In our galaxy, as you go towards the centre, things become much more extreme in almost any way you can describe. The density of stars increases, the speeds of stars increase and the strength of other characteristics, like their magnetic fields, increases. I like to think of it like an urban centre, and we're out here in the suburbs where everything's a little slower and calmer. The centre of the galaxy takes everything to the extreme, basically.

Is that what makes it so hard to study the area at the centre of our galaxy?

The centre of our galaxy has the advantage of being really close compared with the black holes in other

ELENA ZHUKOVA/UNIVERSITY OF CALIFORNIA/UPI/ALAMY STOCK PHOTO

galaxies, so we have some advantages in terms of sorting out what's going on there. The disadvantage is that we're looking through the plane of our own galaxy to perceive what's at the centre.

In addition to having a lot of stars in it, our galaxy also has a lot of dust. That dust makes it difficult for light that's emitted from the centre of the galaxy to reach us. If we were to try to look at the wavelengths that our eyes detect, we would perceive very little, because only one out of every 10 billion of these kinds of photons makes it to us.

So it is less that there is so much going on there, more that there is stuff in between us and there?

Yes, although it is true that in the centre of the galaxy the crowding of stars becomes an issue as well. Of course, that just gets more and more problematic as the galaxy centre becomes further and further away. So our own galaxy is still our best hope for making any detailed measurements. But there are also technical challenges to that. When observing from ground telescopes, the atmosphere blurs the images. The atmosphere is great for us. But it is a total headache in terms of astronomical imaging.

It seems like a lot of our best knowledge about Sagittarius A* comes from just a few stars, including your work. Why is that?

It is true that today there is one star that is, so to speak, the star of the show, called S0-2. It is absolutely my favourite star in the universe.

But we are measuring thousands of stars, and they're all important, they just have different roles to play. Behind those measurements of S0-2, you need stars that tell you how to line up these images across all your observations. So many stars are playing what I would call supporting roles, but they are still absolutely essential.

What is so special about S0-2?

It has a really short orbit, and what I mean by "really short" is shorter than a human lifetime, or maybe a career. It only takes about 16 years for S0-2 to complete an orbit of Sagittarius A*. To put this in context, the sun takes 200 million years to go around the centre of the galaxy. You are not going to wait for that to happen or try to see the curvature of that orbit. It's the orbits of S0-2 and a few other stars like it that give us evidence that there must be a compact, massive object – a black hole – there.

What was it like when you finally got that proof that there was, in fact, a supermassive black hole at the centre of our galaxy?

Oh my goodness, this has been such an exciting project to do because every stage of making progress towards the answer to the question "is there a supermassive black hole?" has been so much fun and so exciting. There's nothing like doing a project where people don't think it's going to work.

›

Speaking of exciting, how has it been since you won the Nobel prize in physics?

It has been surreal. It's surreal to get the Nobel prize, period. It's something that I never anticipated. To get it in the middle of the [covid-19] pandemic adds another element of surrealism to already surreal times. It was really lovely to have good news to share with friends and family and colleagues during these hard times.

All of a sudden, there are a lot of opportunities and invitations to do things, and it forces you to think: what are you going to do now? What are your responsibilities that are associated with receiving a prize like this? What are the opportunities that you want to pursue? I really feel strongly about both taking some of the responsibilities of being a spokesperson for science, but also continuing to pursue the cool questions at the centre of the galaxy.

A lot of women in the sciences, particularly in physics, can feel unwelcome. How can we make the field more accessible and welcoming to everyone?

I think the best thing that you can do is do good science, to show that women can be just as effective at being a scientist as anyone else can. The more women that succeed at the very top, the more I think it helps the field change just through demonstration. And that demonstration is partially for your peers, but probably more importantly, it shows the next generation the possibilities. In my book, the best way you can change the field is by having the people who are in the minority succeeding.

Thinking about black holes, what is the next big question we need to answer?

There are lots. We still don't understand what a black hole is – that is certainly a big question. How do we make quantum mechanics come together with general relativity to explain these objects? I think that is an enormous question that really drives so much of our work. We're still nowhere near answering it. ▮

ESSAY

If black holes existed in reverse, they could give us our first glimpse of the quantum source of space-time, says Carlo Rovelli.

BLACK HOLE... WHITE HOLE?

PROFILE
CARLO ROVELLI

Carlo Rovelli is a physicist at Aix-Marseille University in France and author of books including *Seven Brief Lessons on Physics*, *Reality Is Not What It Seems* and *The Order of Time*

GENERAL relativity predicts that infalling matter ends up concentrating onto a single central point of infinite density, called a singularity. This is a sort of end to reality, a point where time itself stops and everything vanishes into nothingness. But this prediction isn't reliable, because the centre of the hole is outside the domain of Einstein's great theory. Here, gravity is so strong that quantum effects can't be neglected. To understand what happens, we need a quantum theory of gravity.

→ Chapter 6 has more on quantum gravity

Quantum theory has a habit of solving problems of this kind. At the beginning of the 20th century, classical theory predicted that the energy of an electron orbiting the atomic nucleus would spiral down infinitely. Quantum theory clarified why this doesn't happen: it is forbidden by the discreteness of energy. The energy of the electron can change only by specific amounts, and it has a finite bottom level.

Quantum effects can similarly prevent infinite density from forming at the centre of a black hole. In this case, it is the discreteness of space-time itself, predicted by quantum theories of gravity such as loop quantum gravity (which I work on) that does the trick. There are no infinitely small points where density can become infinite. Space is composed of individual units, or quanta, which are small but finite. Falling matter can squeeze into a super-dense state, called a Planck star, but no more. And then? Then matter can do what matter commonly does at the end of a fall: bounce.

It can't bounce up within the black hole, where things can only move downward. But here is the magic. Quantum gravity allows the entire space-time geometry of the black hole to bounce – that is, to continue across the central point of the black hole into a separate and new region of space-time, where not just matter but also the entire space-time is bouncing out. This is what we call a white hole.

A ball that bounces up follows a trajectory that looks like a movie of its fall projected backwards. A white hole is like a movie of a black hole projected backwards. From the outside, it isn't much different: it has mass just like a black hole, so things are attracted by it and can orbit around it. But whereas a black hole is surrounded by a horizon through which it is possible to enter but not to exit, a white hole is surrounded by a horizon through which it is possible to exit but not to enter.

The theoretical possibility of white holes is predicted by general relativity. They are exact solutions of the equations of the theory. But they have long been viewed as mathematical niceties not representing anything real, precisely as were black holes in the past, because it was hard to see how they could originate.

As early as the 1930s, however, Irish physicist John Lighton Synge saw that a minimal adjustment in the solution of the equations of general relativity might allow the possibility that the geometry of the interior of a black hole could continue into a white hole. Quantum mechanics can permit such adjustment.

Where would the daughter white hole be located? Would it be far away, connected by a wormhole, or in a different universe? No, we don't need outlandish speculations. It will be found at the same place where the black hole was, only in its future. Because of the peculiar elasticity of space-time as understood with Einstein's theory, "the other side of the centre" can simply be in the future of the hole. This is hard to visualise, but the result is simple. In the first part of its life, the hole is black and matter falls in – but during the second, after the quantum transition, it is white and matter bounces out.

For this to happen, there should be a moment in which the horizon switches from being that of a black hole to being that of a white hole. Here again, it is quantum theory that allows this to happen, thanks to a well-known phenomenon known as quantum tunnelling. This is a brief violation of the standard, classical equations of physics that can happen, with low probability, even where one wouldn't expect strong quantum phenomena. Tunnelling is what gives rise, for instance, to nuclear radioactivity. A particle trapped inside the atomic nucleus wouldn't be able to escape according to classical mechanics, but quantum theory allows it to "tunnel below" the potential wall that traps it, and thus radiate outside the nucleus.

Tunnelling takes time. Radioactive substances remain quasi-stable for millennia. Similarly, black holes have long lifetimes. If we were to buy the classical theory, a black hole would be eternal. But nothing is eternal. Stephen Hawking showed that quantum theory implies that black holes slowly evaporate and shrink. As they shrink, the probability that they tunnel into a white hole increases. At some point, it happens. And again, the important thing is the geometry of space-time itself – this is the thing that is tunnelling. Instead of evolving according to the equations of classical general relativity, it suddenly tunnels from a black to a white hole horizon.

There is a puzzling aspect to this picture. We see black holes that are millions of years old, therefore a very long time is needed for a large black hole to tunnel into a white hole. But matter falling into the hole reaches the centre rapidly, in a matter of seconds. It would be equally quick to bounce out again. How can matter then find itself exiting a white hole so soon, when forming a white hole takes so long?

The answer is enchanting. Time is incredibly flexible in general relativity. We know it passes more slowly at sea level – closer to Earth's centre – than on the mountains. Approaching a massive star or a black hole, it slows down even more. And this solves the puzzle: a very short time inside the hole can match a very long exterior time. Seen from the outside, the internal evolution of the hole appears like a bounce, but in super slow motion. The holes we see in the sky may simply be objects that collapse and bounce back out, perceived by us from the exterior in exaggerated slow motion.

A bonus from this scenario is that it solves the famous black hole information paradox – we expect information never to be completely lost in nature, but it is lost if time comes to an end inside a black hole. The solution is simple: if anything ends up bouncing out, information is recovered.

↑
See Paul Davies's essay on page 57 for the lowdown on the information paradox

All this gives an appealing scenario for the full life evolution of a black hole. In the interior of the hole, there is no singularity, no place where space-time ends, and seen from the exterior a black hole isn't eternal. Rather, at some time, the black hole turns white and whatever fell inside it escapes.

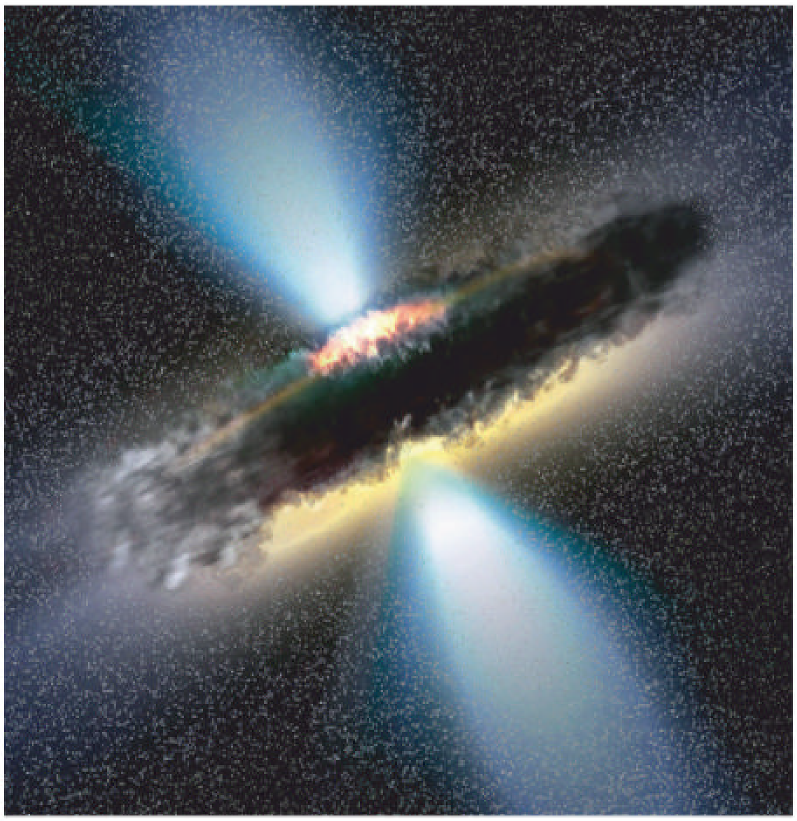
The scenario is theoretically beautiful. Does it imply that the sky is truly full of white holes? And if so, can we see them?

The answer depends on things we don't yet fully understand. Most of the black holes we see in the sky are formed by the collapse of a star. These are too young and big to have already tunnelled into white holes – big holes live longer. But it is possible that smaller black holes formed in the fierce environment of the early universe, shortly after the big bang. These primordial black holes might have already tunnelled into white holes, or may be tunnelling today. But we aren't sure about their number, and this makes predictions about current white holes uncertain.

A further source of uncertainty is the lifetime of a black hole. Detailed calculations have been attempted using loop quantum gravity, but they depend on approximations and aren't yet conclusive. Still, we have a pretty firm bound between a maximum "long" lifetime, limited by Hawking's evaporation time, and a minimum "short" lifetime required by the onset of quantum phenomena. This allows us to draw some preliminary conclusions.

If the lifetime is long, only small primordial holes have already turned white. This would mean most of the white holes currently in the sky should be of minimal size. The minimal white hole size is Planckian, namely around a microgram, or the weight of half an inch of a single strand of human hair. This is an intriguing possibility because white holes of this size can be relatively stable, and they could be a component of the mysterious dark matter that astronomers have

If we could spot a black hole turning white, it would be a glimpse of quantum gravity in action


ESA/V. BECKMANN (NASA-GSFC)

(indirectly) detected in the sky. Most other hypotheses on the nature of dark matter demand modifications of well-established laws of physics. For instance, they rely on theories predicting new entities called super-symmetric particles. But failures to detect such particles have raised questions about these theories.

←
Turn back to page 39 for more on dark matter

The possibility that dark matter is instead composed of small black holes doesn't require anything more than established physics, namely general relativity and quantum theory, and isn't ruled out by any observation. If this is correct, we have already observed white holes: they are the dark matter!

Alternatively, if the black hole lifetime is short, primordial black holes tunnelling today should have the mass of a small planet and could be exploding violently, transforming most of their mass into emitted radiation. This event should send us extremely energetic cosmic rays and short, violent signals in the microwave or radio band. The latter are particularly intriguing because similar signals have already been detected: the mysterious fast radio bursts recently observed by radio telescopes. Again, we might have already seen white holes.

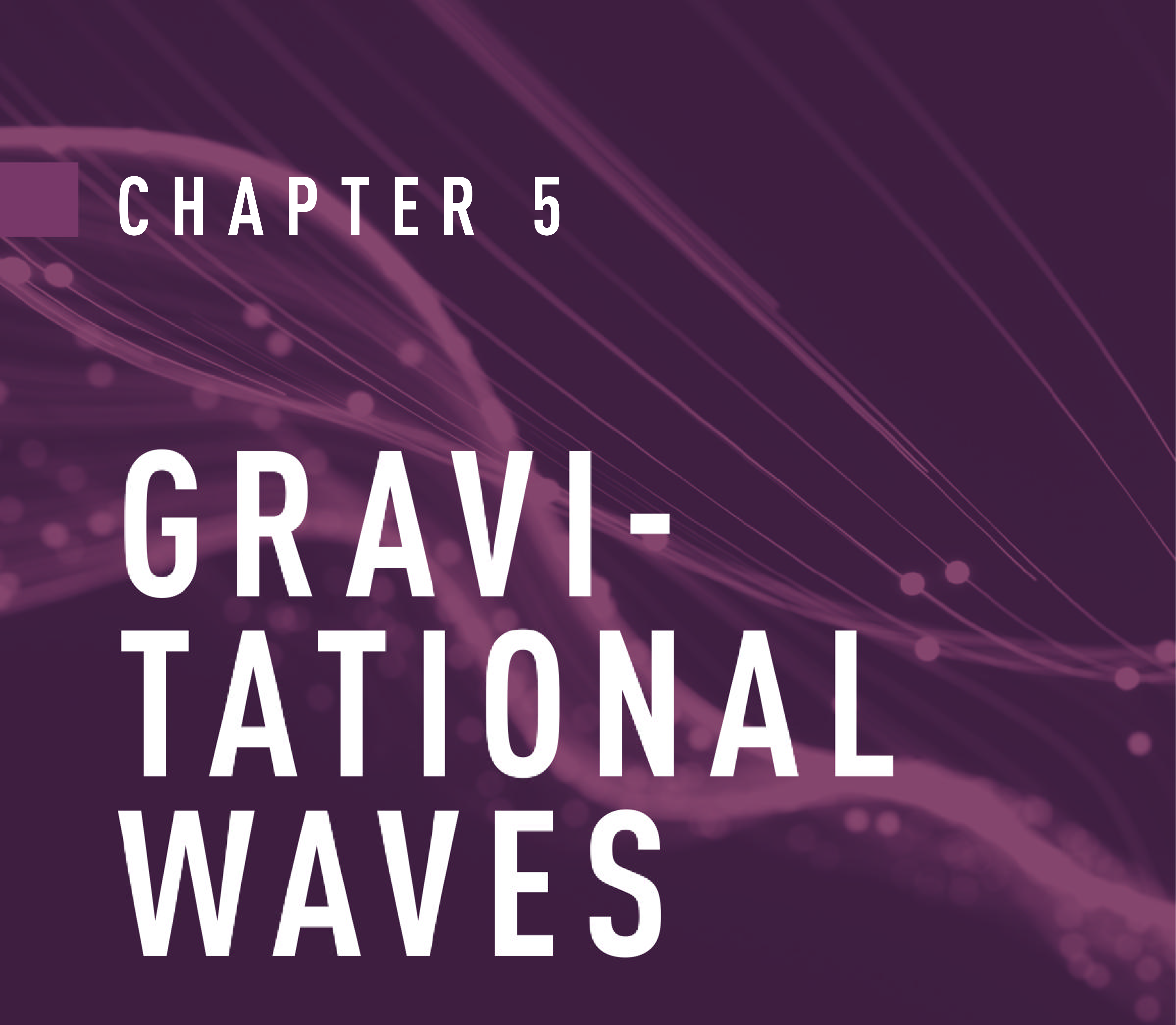
We can't confirm that these signals are indeed from white holes with just a few detections – other sources are possible. But there is a signature we will look for across a large sample: a flattened redshift. Signals emitted by distant and therefore younger white holes should produce shorter wavelengths than nearby, older ones. This is something we might be able to spot in high-energy cosmic rays or fast radio bursts once we have enough data. If we do, we will have some evidence that white holes exist. Finding evidence of white holes in the sky would be a beautiful step ahead in our understanding of the universe. They could represent the first direct observation of quantum gravity at work, and so open a window on the greatest problem in fundamental physics, the problem of understanding the quantum aspects of space-time.

I close by mentioning one last very speculative idea. Our universe might not have been born at the big bang: it may have bounced out from a previous collapsing phase. This possibility is allowed by loop quantum gravity and other theoretical frameworks. The quantum mechanism of the cosmic bounce is similar to the black-to-white-hole bounce. The Planckian white holes in today's dark matter could have formed before the bounce. If so, the geometry of space-time at the bounce wasn't homogeneous as conventional cosmology suggests, but rather very crumpled, because each white hole is like a long spike out in the geometry of space-time.

→
Turn to page 88 for more on bouncing universes

This fact could be relevant for the mystery of the arrow of time – the question of why time goes in only one direction. The arrow of time might not be caused by the initial state of the universe being "special" (that is, low entropy), as is commonly believed. Instead, it may be a perspectival phenomenon related to the very "special" location of us observers: we are outside all the holes.

White holes are a plausible – albeit almost completely unexplored – possibility. We are yet to identify one, but then we didn't recognise black holes for long time either. ∎



CHAPTER 5

GRAVITATIONAL WAVES

Albert Einstein was a prolific predictor of new phenomena – and often had difficulty coming to terms with the results. So it was with black holes, and so it was with gravitational waves.

These minuscule ripples in the fabric of space-time are given out when massive objects move in the cosmos, for example merging black holes. For a long time, they remained the only experimental prediction of general relativity not to be verified.

That all changed with a stunning detection in 2015, almost exactly 100 years after Einstein first presented general relativity to the world. It opened up a new era in astronomy.

THE LONG ROAD TO DETECTION

Gravitational waves eluded detection for a century after Einstein predicted them. That had two main causes – scepticism on the part of theorists, and the sheer scale of the challenge faced by experimentalists.

PREVIOUS PAGE: JUST_SUPER/ISTOCK
SPANTELDOTRU/ISTOCK

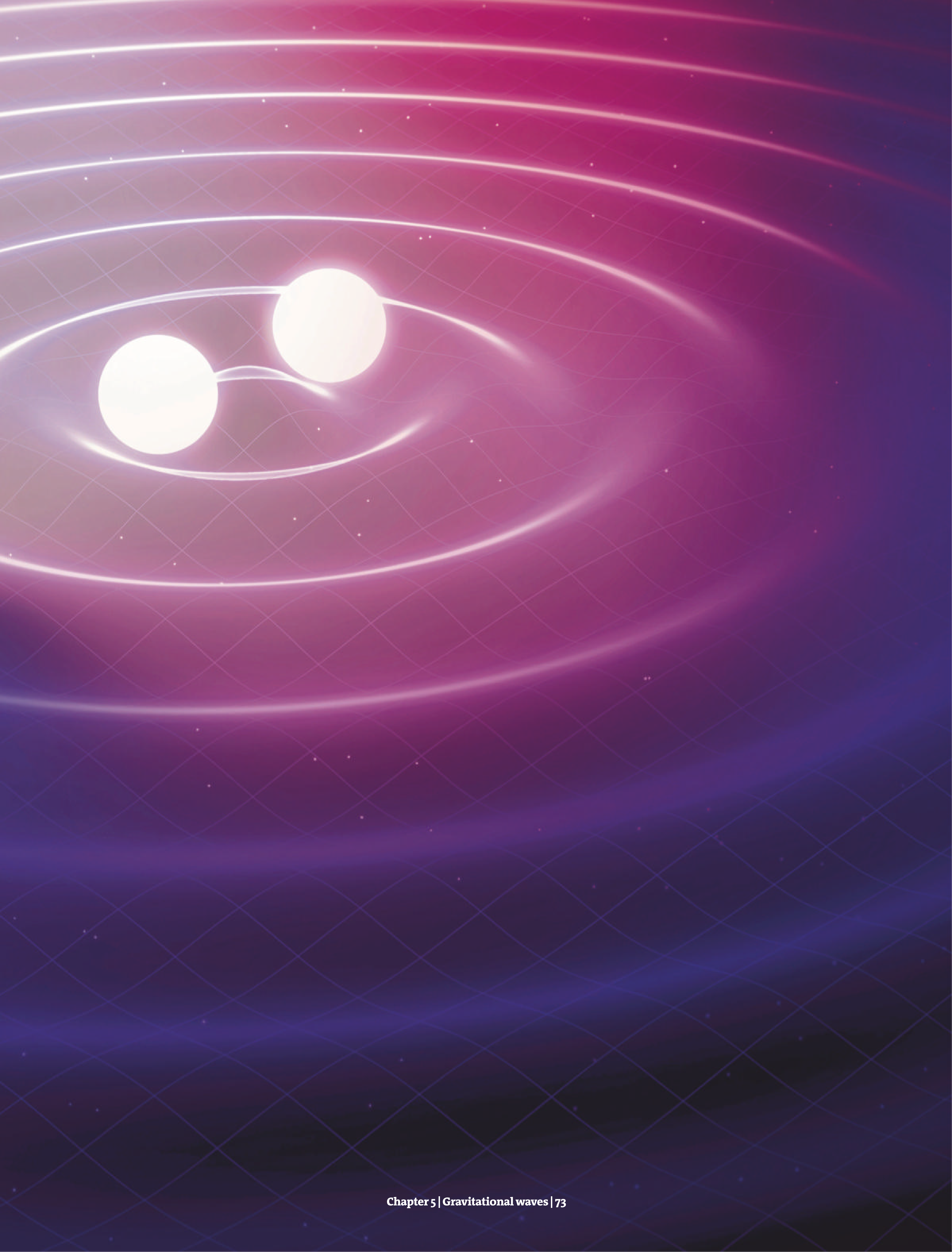
MATTER causes space-time to curve, says general relativity. When a large mass accelerates, that curvature should change. The result is ripples in space-time that spread out at the speed of light, just as electromagnetic waves generated by accelerating electric charges spread.

Gravitational waves are actually a subtle consequence of the special theory of relativity derived by Einstein in 1905, where they serve to prevent gravitational influences propagating across space-time instantaneously, faster than the speed of light.

The analogy with electromagnetic waves, also bound to the speed of light, isn't perfect: rather than propagating through space-time as electromagnetic waves do, gravitational waves are contractions and expansions of space-time itself. Because gravity is much weaker than electromagnetism, they are also minuscule by comparison.

All of that makes calculations with gravitational waves rather hairy. Einstein realised straight away that the equations of general relativity had wave-like solutions, and in 1918 he derived a formula that allowed him to estimate how much energy these waves should carry. But general relativity's core equations are so intractable that controversy persisted for decades over whether the formula was even theoretically sound.

Almost two decades later, together with a young colleague, Nathan Rosen, Einstein titled another paper sent to *Physical Review* "Do gravitational waves exist?", answering the question with a cautious "no". ›



The paper never appeared: the journal's editor sent it on to an anonymous referee. Offended, Einstein withdrew it. He didn't take kindly to being peer-reviewed.

But that gave him time to realise the mistake in his calculations and reverse his prediction before the paper was eventually published in 1937 in the *Journal of the Franklin Institute*. The existence of ripples in the fabric of reality, extending every which way throughout the cosmos, finally had Einstein's personal seal of approval.

It was only after Einstein's death that gravitational waves became widely accepted, however. Experimentalists duly built detectors – initially large, suspended cylinders that might be nudged by a passing wave. In the late 1960s, US physicist Joseph Weber was the first to claim a sighting. More than a dozen such claims followed, but none stood up to scrutiny.

But even before we had detected them directly, we already had overwhelming indirect proof that they exist. In 1974, astronomers Russell Hulse and Joseph Taylor discovered a binary pulsar – an orbiting pair of neutron stars beaming out radio waves at precise intervals – and started tracking its rotation rate. By the early 1990s, they had shown that the stars were losing energy at precisely the rate Einstein predicted they would if they were emitting gravitational waves. A handful of similar binary systems studied since then have confirmed this view.

There is one big reason why a direct detection took so long. "Space-time is really, really stiff," says Vernon Sandberg at the Laser Interferometer Gravitational-Wave Observatory (LIGO) collaboration that eventually made the breakthrough. It is something like a thousand billion billion times stiffer than diamond, so even mighty cosmic events generate pitifully weak, murmuring waves in it. A gravitational wave from the death tango of two neutron stars that each had a mass of 1.4 suns – the benchmark used by gravitational-wave astronomers – would squeeze or stretch the 4.3 light years between us and the star system Alpha Centauri by less than half the width of a human hair.

↓
The next section has more on the LIGO detector

This made the announcement on 11 February 2016 by researchers at LIGO that they had finally spotted a gravitational wave all the more sensational. The signal was actually picked up by LIGO's two observatories in Hanford in Washington and Livingston in Louisiana on 14 September 2015. The five-month delay was down to painstaking analysis and cross-checking to ensure that the signal couldn't have been from any other source.

It wasn't. It fit perfectly with the sort of signal we would expect if two black holes, each about 30 times the mass of the sun, collided. The details of the signal suggest that they circled each other closer and closer until they finally merged into one.

This immediately resolved one open question for astronomers. Before the signal came in, the very existence of such black hole binaries was contested. Because they are dark, black holes of these masses are almost impossible to spot unless something bright – like a star – orbits them.

←
Turn back to chapter 4 for more on black holes

Today, LIGO – now working in consort with a third gravitational wave detector, Virgo, located in Italy – has many tens of confirmed detections under its belt. These include gravitational waves from the death spiral of two neutron stars. Unlike black holes, which hide their mass behind an event horizon even as they crash, colliding neutron stars spew hot, bright matter across space, which could help us explore other mysteries. Studying these explosions may explain short gamma-ray bursts, mysterious and incredibly bright electromagnetic phenomena. They might also help explain where much of the universe's heavy elements, like uranium, thorium and gold, are forged. ›

THE GRAVITATIONAL WAVE BACKGROUND

Everything in the universe is constantly being stretched and squeezed by gravitational waves. That suggests a way we can make detections of a background of waves permeating the universe using normal radio telescopes, rather than making single detections with ultra-sensitive detectors on Earth or in orbit.

That is the guiding idea behind the North American Nanohertz Observatory for Gravitational Waves (NANOGrav) consortium, which uses a so-called pulsar timing array to attempt to build a sort of map of gravitational waves. Pulsars are neutron stars that rotate extremely rapidly and regularly, sending out beams of light that act as "ticks" in extraordinarily precise cosmic clocks.

When a gravitational wave passes through the same region of space-time that those beams of light are travelling through, it makes the light appear to take slightly more or less time to reach us, causing the "ticks" from a pulsar to seem irregular. Pulsar timing arrays require radio telescopes to observe the signals from many pulsars across the sky simultaneously.

"These pulsars are spinning with millisecond periods and we are able to detect changes in the time of arrival [of signals]… at the hundreds of nanosecond level," says Joe Simon at the University of Colorado Boulder.

Early in 2021, Simon and other NANOGrav researchers presented data gathered on 45 pulsars over the course of 13 years, and found a gravitational wave signal that was identical across multiple pulsars. This strange, low-frequency hum could be the first evidence of the gravitational wave background. To prove that this signal is from the gravitational wave background, however, they would need to see a distinctive pattern in the gravitational waves affecting each pulsar.

Gathering the additional data necessary to find that pattern should only take about a year, says Simon, although analysing it may take longer.

If the signal is in fact the gravitational wave background, it will be a useful tool for understanding the most massive objects in the universe. "This will tell us more about black holes in the universe, and especially the supermassive black holes in galactic centres," says Nelson Christensen at the Observatory of Nice in France. "This NANOGrav signal is likely from [black hole] binaries with billions of solar masses." As these enormous pairs of black holes merge, they emit thrums of gravitational waves powerful enough to persist throughout space-time.

This effort will build a bridge between the gravitational waves we have already spotted coming from smaller black holes with the Laser Interferometer Gravitational-Wave Observatory (LIGO) and Virgo detectors, and those from supermassive black holes, says Christensen. Such a bridge will help us understand how different types of black holes form, how galaxies evolve with the black holes within them and maybe even the larger, mysterious forces at work in our universe like dark matter and dark energy.

Such single event detections are just the start, however. Put several together and we should be able to get new insights into the history and composition of the universe as a whole, says Avi Loeb at Harvard University. The signals from several black hole mergers, for example, can be combined to help understand the nature of dark energy, which is causing the universe's expansion to accelerate.

From the "shape" of the signal – how the waves' frequency and volume rise and fall – we can discern the sizes of the black holes involved, and determine how loud the event was at its source. Comparing how powerful it really was to the faint vibrations that LIGO detected tells us how far away it occurred. Combined with observations from standard telescopes, this can show us how space has expanded during the time that the waves took to reach us, providing a measure of dark energy's effect on space.

←
Turn back to page 43 for more on dark energy

This measure should be stronger and more reliable than anything we have used so far. "If you have tens of [detections], it will be a new branch in cosmology," says Loeb.

Other researchers hope to use gravitational wave signals to put Einstein's general theory of relativity to even more stringent tests. One way is through the equivalence principle, an assumption that gravity affects all masses in the same way. "In the age of GPS and space travel, where even minute deviations from the assumed theory of gravity would have major consequences, it is of enormous importance," says Xue-Feng Wu at Purple Mountain Observatory in Nanjing, China.

Erminia Calabrese, an astronomer at Cardiff University in the UK, sees gravitational waves as a way to check whether gravity behaves as relativity predicts it should over large distances. "If their strength fell off with distance in a surprising way, we could detect this with the upcoming LIGO data," she says.

New detectors are already in the pipeline. The European Space Agency is working on the Laser Interferometer Space Antenna (LISA), a huge, space-based detector due to blast off in the mid-2030s. A preparatory proof-of-principle experiment, LISA Pathfinder, was in orbit from 2015 to 2017 and was deemed a success.

Further ahead, we might see more sensitive gravitational wave detectors, working at shorter wavelengths than LIGO. These may allow us to sense primordial gravitational waves from the very young universe. These waves should have been produced in the period of inflation – the tremendous growth spurt in the first instants after the big bang. Unlike photons and other electromagnetic radiation, they would have travelled freely through the newborn universe. At the moment, we can only see as far back as 380,000 years after the big bang, when the universe became transparent to light and the cosmic microwave background was emitted.

←
Chapter 2 has more on the cosmic microwave background and inflation

Gravitational waves may even point the way towards a unified theory of the universe. Theory suggests that at some point in the universe's history, all four fundamental forces were united into a single force. As the universe expanded and cooled, the forces split off from one another in a series of as-yet poorly understood events. Another suggestion is that jitters in gravitational wave signals could reveal the existence of gravitons, the hypothetical quantum particles of gravity, and so point to a theory beyond general relativity that unites all four forces on the basis of quantum theory.

→
Chapter 6 has more on quantum gravity

Now, Maulik Parikh at Arizona State University and his colleagues have calculated that the existence of gravitons could create jitters in gravitational wave signals. They found that these could theoretically be detected with current gravitational wave observatories.

All of that goes some way to explaining the buzz surrounding gravitational waves. "There's going to be a revolution", in the words of LIGO team member Erik Katsavounidis at the Massachusetts Institute of Technology, in response to the first discovery. ▮

ENRICO SACCHETTI

HOW WE FOUND GRAVITATIONAL WAVES

It took decades to prove that gravitational waves are real, as the people who made it their life's work recall

In 1969, ***Rainer Weiss*** *was a young professor at the Massachusetts Institute of Technology (MIT). At the time, gravitational waves were a theoretical curiosity: Einstein himself took years to be convinced by his own prediction that moving cosmic bodies would send out ripples through space-time. Then physicist Joseph Weber claimed to have recorded one on a xylophone-like instrument he called a resonant bar detector. Weiss takes up the story.*

The students on my course were fascinated by the idea that gravitational waves might exist. I didn't know much about them at all, and for the life of me I couldn't understand how a bar interacts with a gravitational wave.

I kept thinking, well, there's one way I can explain how gravitational waves interact with matter. I said, suppose you take a light – I was thinking of just light bulbs because, in those days, lasers weren't yet really there – and sent a light pulse between two masses. Then you do the same when there is a gravitational wave.

Lo and behold, you see that the time it takes light to go from one mass to the other changes because of the wave. If the wave is getting bigger, it causes the time to grow a little bit. If the wave is trying to contract, it reduces it a little bit. So you

›

can see this oscillation in time on the clock.

I was hiding in this little office in Building 20 at MIT, and for about three months I thought about how you might do this. First, I thought you couldn't get clocks good enough. But we did some experiments and I learned you could do unbelievably exquisite measurements with lasers.

I wrote this up, but I didn't publish it. The people at MIT wanted to know where the hell I'd been spending my time, so I put it into the quarterly progress report for my lab. I came to the conclusion that, if you built this thing big enough, you could probably detect gravitational waves.

Based on Weiss's idea, the US National Science Foundation (NSF) eventually began funding the development of what became the Laser Interferometer Gravitational-Wave Observatory (LIGO) in 1979. But progress was slow, and when physicist ***Barry Barish*** *at the California Institute of Technology took the lead on the project in 1994, questions were being asked.*

Previous page: The concrete housing of LIGO's two beam tubes stretch off die-straight into the distance

Above: Within the beam tubes, lasers ping back and forth hundreds of times, reflected by mirrors

The NSF had lost confidence and were basically giving up on it. There was a lot of resistance. It represented the extreme of what you might call a high-risk, high-payoff project. So we revamped it entirely. Over a period of six months, we made it look like a new project. And it was hard, because if you had asked me then, could we build what we now know we need to detect gravitational waves, the answer would have been no.

So the idea, and the words that I used, was that with the initial version of LIGO, it would be possible to detect gravitational waves. And then we would evolve into a detector, which we called Advanced LIGO, where it would be probable. But in all honesty, we had nothing more

than ideas on how to do the Advanced LIGO part. To me, the miracle in this whole thing is that we somehow got financial support for 22 years until we succeeded.

*LIGO's detectors occupy two sites in Livingston, Louisiana, and Hanford, Washington. Since joining the project in 2000, Hanford's lead detection scientist **Michael Landry** has been ensuring that the instrument is as sensitive as possible to the minuscule signals it was built to detect.*

Space is a stiff medium, so it doesn't want to vibrate. The detector has to register changes that are about a thousandth the size of a proton. If you were trying to measure the distance between here and our nearest star, Proxima Centauri, it would be like watching that distance change by the width of a human hair.

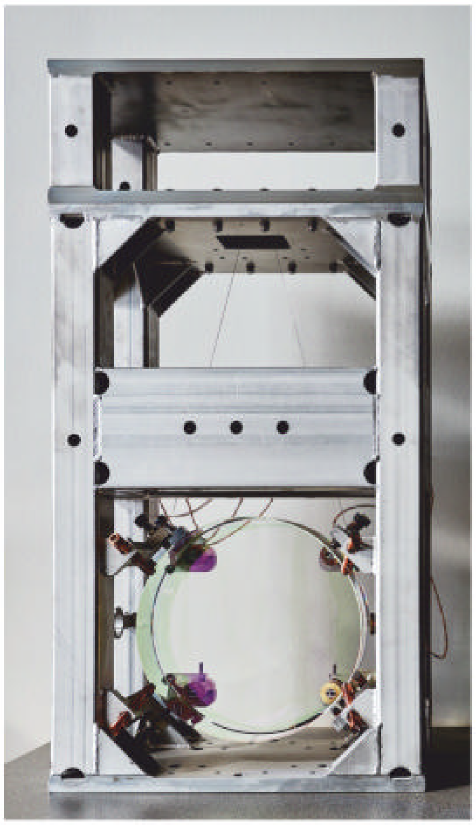
It is an ongoing battle to suppress noise in the instrument. There are ground noises like earthquakes, but a less obvious example is Earth's ringing – at low frequencies, it rings like a bell because of ocean waves hitting against the continental shelf.

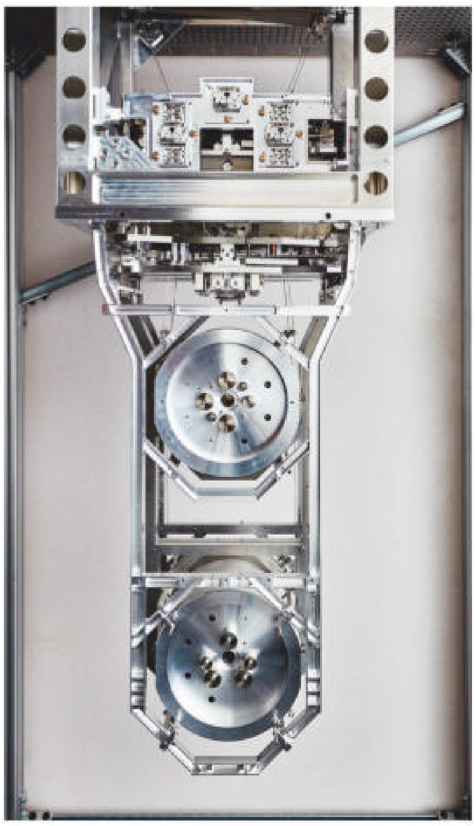
The "vertex" where the two laser beams are recombined. The interference pattern they create changes when a gravitational wave disturbs either one of them

If you get storms off the coast of Alaska or the Gulf of Mexico, ground motions increase. We have to suppress that motion by registering it with seismometers and feeding it into seismic suppression systems – kind of the way noise-cancelling headphones work by sampling the ambient noise and then playing it with the right phase to cancel it out at your ear.

There are also a lot of internal noises that we have to suppress. Things like electronic noise, or quantum noise in the laser. All of that means the LIGO detectors are absolutely the most quiet and the most sensitive detectors ever built.

›





Far left: LIGO's cylindrical mirrors are polished to nanometre precision to reflect the laser beams without absorption and ensure no surface roughness affects the detector's timing

Near left: The ultra-pure silica mirrors are suspended on glass fibres. Their suspension cage adjusts for the changing motions of the ground that might otherwise disturb the measurement

Right: The LIGO Livingston control room in Louisiana, where a distinctive "chirp" was recorded at 09:51 UTC on 14 September 2015. It lasted just a fraction of a second, but was registered almost simultaneously 3000 kilometres away in Hanford, Washington – the expected signal of a gravitational wave travelling at light speed

In September 2015, just a few days after Advanced LIGO finally came on stream, ***Michael Landry*** *received notice of an unexplained signal. At first, he was convinced it was an "injection" – an artificial pulse used occasionally to test the instrument.*

On the morning of 14 September, I opened my computer and saw an email indicating this event seen in the data literally tens of minutes earlier. I thought, it's probably an injection. It was so early in the observation phase. It wasn't until later, in the lab, that we determined there was no such injection. We did a whole lot more investigation and it took months to validate it. But it was immediately obvious that if this thing wasn't an injection, it was the best damn thing we had ever seen.

The gravitational wave apparently came from the collision of two black holes, formed from collapsed stars, 1.3 billion light years away. For ***Rainer Weiss****, 83 at the time and a professor emeritus at MIT, it was a long-sought vindication.*

The discovery itself was spectacular. To me, it was the thing that I would have wanted most, to see the collision of two black holes. If you want to ask, what was the reason for building this thing in the first place, it's to check up on whether Einstein's theory works in strong gravitational fields. That was the one place where general relativity hadn't been tested. And here, suddenly, we have in our hands a thing that says Einstein's field equations, the whole thing, is absolutely right.

For ***Nergis Mavalvala****, an MIT physicist who had worked on LIGO for 25 years, this detection was just the beginning.*

I don't think this was some master plan from nature, as in "let's be nice to these people here on this little Earth place". Black holes collide all the time. What was lucky was that we happened to have a detector with sufficient sensitivity to see this one at the moment it went up. If we continue, we will see more.

One of the amazing things about general relativity is you solve the equations – although that's taken decades – and create templates for how signals should look. Nature was kind in that the very first signal we saw was so clear. Many people expected we would see really weak signals, barely poking above the noise, and that there would be a lot of discussion about whether it was a detection or not. None of that happened. That's where we lucked out.

The discovery drives us harder because we know there is stuff out there waiting to be observed. You might imagine we would think, "OK, now we've seen it we can pack up and go home". But in fact it's just the opposite. We've seen the very first gravitational wave, but we have so much more to discover.

We have a lot to learn about black holes, and then there are neutron stars. Personally, my hope is that we will see something that really has us scratching our heads. Maybe we will have discovered some new object that I can't begin to describe or name. ▮

The pictures of LIGO accompanying this essay are part of a specially commissioned set by Enrico Sacchetti

HOW DOES LIGO WORK?

Each of LIGO's detectors, at two sites in Washington and Louisiana, consists of two 4-kilometre-long arms. They are identical in almost every respect. The arms each contain a steel pipe housing a vacuum through which one half of a split laser beam pings back and forth, bouncing off a mirror at the far end. Back at the main building, this reflected laser beam is reunited with its twin that has zipped up and down the other arm.

Any change in the length of either arm caused by a passing gravitational wave distorting space-time would alter the patterns of constructive and destructive interference that form when the two light waves recombine. Gravitational waves squeeze space in one direction while stretching it in the other, making the LIGO detector with its arms splayed at right angles doubly sensitive: one arm of the interferometer should lengthen and then shorten as a gravitational wave passes, while the other shortens and lengthens. A standard-issue gravitational wave is expected to expand and contract LIGO's arms by just 10^{-19} metres, a distance one ten-thousandth the width of a proton.

Hence why LIGO has two twin detectors 3000 kilometres apart. See the same signature in one and 10 milliseconds later in the other (the time it would take a wave to travel between the two) and then you would be talking.

Or perhaps not: to make that delicate a measurement, every possible disturbance, every tiniest reverberation that might put things out of kilter must be suppressed. To increase the detector's sensitivity, light is made to bounce back and forth many times along the interferometer arms. In the upgraded "Advanced LIGO" configuration that made the first detection, the aim was for 100 reflections, making each 4-kilometre arm effectively 400 kilometres long. For that to work, the reflecting mirrors had be held unerringly still, hung on wires of fused silica glass barely a millimetre thick to isolate them from motions of the ground as far as possible.

Even with such precautions, the signals LIGO detects last just seconds. The gravitational waveform produced by black holes as they spiral towards each other and finally merge lasts for many millions, perhaps even billions of years. That creates a pattern in which both frequency and amplitude increase as the black holes approach each other, orbiting ever faster – but it is only in the final second that the signal reaches high enough frequencies and high enough amplitudes for LIGO to detect the vibration, above the general background noise from other non-cosmic sources.

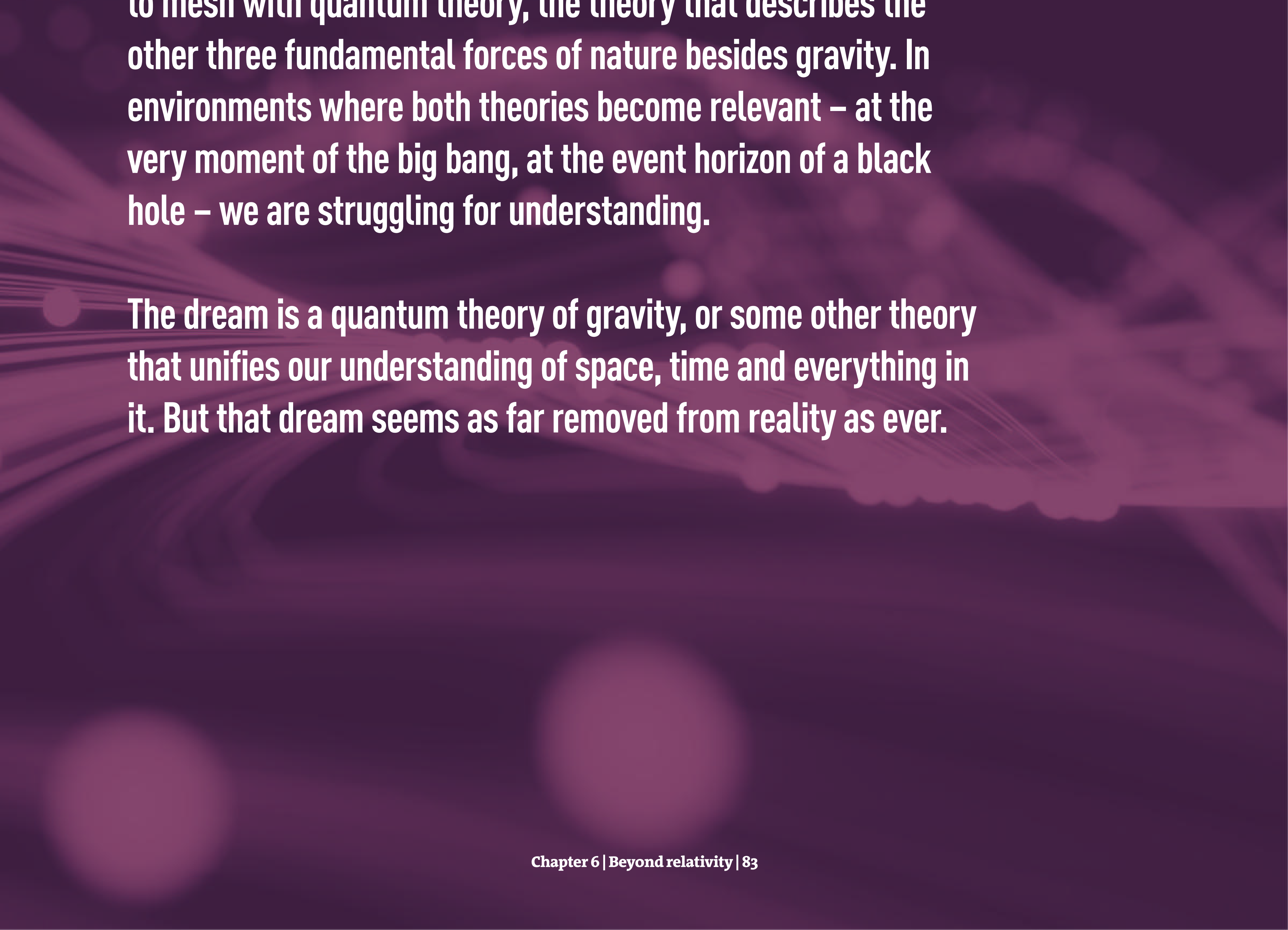
CHAPTER 6

BEYOND RELATIVITY

Albert Einstein's relativity has proved a peerless guide to the universe – as far it goes. But it has limitations and deficiencies that leave physicists and cosmologists aspiring for more.

That becomes most obvious in general relativity's failure to mesh with quantum theory, the theory that describes the other three fundamental forces of nature besides gravity. In environments where both theories become relevant – at the very moment of the big bang, at the event horizon of a black hole – we are struggling for understanding.

The dream is a quantum theory of gravity, or some other theory that unifies our understanding of space, time and everything in it. But that dream seems as far removed from reality as ever.



THEORIES OF EVERYTHING

Since the times of the ancient Greeks, we have had the idea that, if we dig down, we can uncover the ultimate building blocks of reality and the universal rules that govern its behaviour. Today, things don't look quite that simple, not least because of how general relativity fails to mesh with the theory that governs the rest of physics: quantum theory.

PREVIOUS PAGE: JUST-SUPER/ISTOCK
KASEZO/ISTOCK

A RATHER glib distinction is often made between the two pillars of modern physics. Quantum mechanics is the physics of the very small, while general relativity is the physics of the very large. That's not quite accurate – for example, quantum-mechanical effects have been observed spanning hundreds of kilometres. And at some scale, surely these two supremely accurate theories must come together.

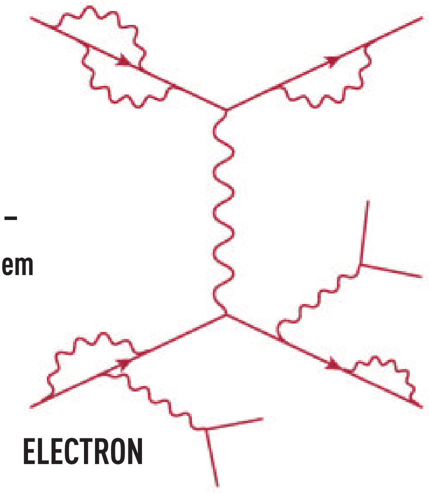
Yet wherever they do cross paths, the two theories fail to play nicely together – such as around black holes. Even things as basic as space and time highlight how badly these two theories get along. Relativity's space-time is a smooth four-dimensional sheet; the quantum field theories that underlie the standard model suggest that space is pixelated into units with sizes of about 10^{-35} metres, and don't even treat time as a real and observable thing.

Efforts to establish a quantum theory of gravity have stumped many physicists over the past century. Asked to choose between the two theories, most would place their money on quantum theory being "right", because its mathematics is such a successful prism through which to view the world. Others, from Einstein onwards, have taken issue with quantum theory's seeming "irreality" and spooky, counter-intuitive correlations between apparently unrelated objects. If we can't find a convincing physical reason why these correlations are just so, they argue, perhaps quantum theory is just an approximation to something better.

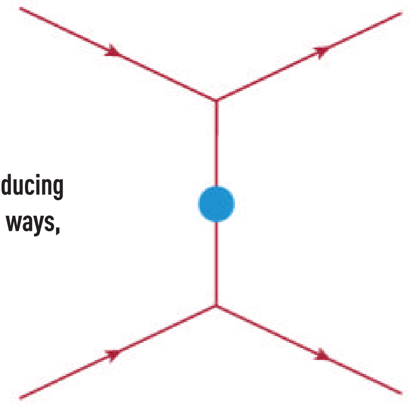
Einstein was just one of those burned by the search, becoming extremely unproductive in his later years as he sought a theory of everything. To understand the difficulties involved, we must start with a fundamental tenet of quantum physics. Heisenberg's uncertainty principle embodies the fuzziness of the quantum world. It allows particles, such as electrons or photons of light, the equivalent of an interest-free loan: they may borrow energy from empty space and use it to make mass, according to Einstein's $E = mc^2$. This mass takes the form of short-lived "virtual" particles. The

›

Gravitons are conjectured quantum particles of gravity – but theories incorporating them tend to be unruly



Particles such as electrons can interact by producing and exchanging massless photons in countless ways, often resulting in infinities in the calculations



only caveat is that they must pay this energy back – the particles must disappear once again – before anyone asks any questions. The more energy they borrow, the quicker this must happen.

Given such freedom, one can imagine an electron, photon or any other particle going to town, taking out many zero-interest loans in succession. As a result, calculating even a prosaic quantum process – an electron travelling from left to right, say –becomes enormously complex. In the words of physicist Richard Feynman, we must "sum over all possible histories", taking into account the infinite variety of ways virtual particles can be produced (see diagram).

The history of applying quantum theory to nature's forces is a history of getting to grips with these unruly infinities. One huge success story is electroweak theory, the theory that combines the electromagnetic and weak nuclear forces to explain how electrons and photons work. Its predictions, of everything from particle masses to their decay rates, are accurate up to 10 decimal places. The eventually successful variant, a bedrock of today's "standard model" of particle physics, tamed the mathematics using a bunch of undiscovered massive particles, the W, Z and Higgs bosons – theoretical predictions that have all since been discovered to be components of reality.

But follow this same principle, and postulate that gravity is made of quantum particles called gravitons, and you hit a problem. Eliminating the infinities that emerge requires inventing a second particle with a mass 10 billion billion times that of a proton. As ever, the larger the amount of energy borrowed, the more quickly it must be paid back, so these fixer particles are very short-lived. This means they can't get very far, and so occupy only a minute amount of space.

But general relativity says that mass bends space-time. Concentrate enough mass into a small area and a black hole will form, a point of infinite curvature in space-time. And this is exactly the guise our new particle takes. Nature plays a cruel joke on us: our scheme to eliminate one sort of infinity creates another.

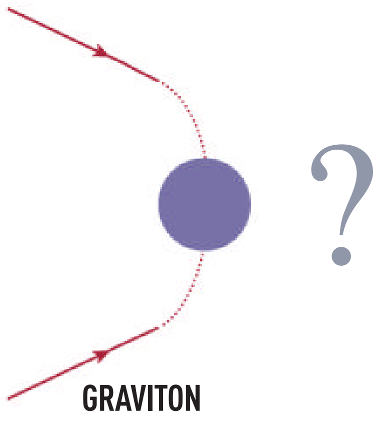
Attempts to get round this fundamental roadblock have led us to destinations such as string theory, which assumes that all particles are manifestations of more fundamental vibrating strings. When we start summing over all possible histories of these "fluffier" objects, the hard infinities produced by virtual particles drop away almost by magic. Another commonly considered idea is loop quantum gravity, which suggests that space-time itself is chopped up into discrete blocks. This pixelation imposes an upper limit on the amount of energy any particle can borrow, again rendering calculations finite.

Despite their seemingly radical assumptions, these two candidate unified theories are in many ways the most conservative extensions of current models: both attempt to preserve as much of the theoretical underpinnings of quantum mechanics and general relativity as possible (see "Six routes to a theory of everything", right).

What about more esoteric ideas, such as changing the rules of the existing game? For instance, if general relativity were to treat space and time separately again, rather than lumping them into one combined space-time, that might provide some wiggle room. But relativity and quantum mechanics both tally so well with reality in their respective spheres that it is devilishly difficult to formulate such tweaks.

That failing means the idea of a theory of everything has quietly disappeared, says Renate Loll at Radboud University in Nijmegen in the Netherlands. "For a while, you would see it in papers, in the heyday of string theory, but it has gone totally out of fashion." Chris Isham at Imperial College London goes further. A theory of everything is "psychologically compelling", he says, but there is no reason to think one exists – or that we can find it. That we have got so far with mathematics is a remarkable fact, but it doesn't mean we can go all the way. Others, however, beg to differ. ▮

The situation is saved by the existence of heavier particles – the W, Z and Higgs bosons that aren't so easily produced, cancelling out the infinities



Performing the same trick with graviton interactions requires a particle so massive that it acts like a black hole – and all calculations are off again

SIX ROUTES TO A THEORY OF EVERYTHING

STRING THEORY

String theory predicts that space has hidden extra dimensions, invoking symmetries embedded in these 10 (mainly tiny) dimensions to "fold" energy into geometric shapes that look like certain fundamental particles, or mimic the way that space curves in the presence of mass. Different string vibrations produce particles with different characteristics: electrons, quarks, the Higgs boson, the hypothetical "graviton".

M-THEORY

A more general 11-dimensional theory that ties together variants of string theory. A puzzling feature is that it predicts many (possibly infinitely many) ways of curling up these dimensions, leading to a multiverse of different universes. Some may look like ours, with three generations of quarks and leptons and four forces; many will not. But from a theoretical point of view, they all seem plausible.

LOOP QUANTUM GRAVITY

It hasn't had quite the same exposure as string theory or M-theory, but so far loop quantum gravity is their only real rival as a theory of everything. The basic idea is that space isn't continuous, as we usually think, but is instead broken up into tiny chunks 10^{-35} metres across. These are then connected by links to make the space we experience. When these links are tangled up into braids and knots, they produce elementary particles. As with string theory, a true experimental test is still some way off.

CAUSAL DYNAMICAL TRIANGULATION (CDT)

Just as loop quantum gravity breaks up space into tiny "building blocks", CDT assumes that space-time is split into tiny building blocks – this time, four-dimensional chunks called pentachorons.

The pentachorons can then be glued together to produce a large-scale universe – which turns out to have three space dimensions and one time dimension, just as the real one does. CDT as it currently stands, however, can't explain the existence of matter.

QUANTUM GRAPHITY

All the theories above assume that space and time exist, then try to build up the rest of the universe. Quantum graphity – the brainchild of Fotini Markopoulou at the Perimeter Institute for Theoretical Physics in Waterloo, Canada, and colleagues – tries to do away with them.

When the universe formed in the big bang, says Markopoulou, there was no such thing as space as we know it. Instead, there was an abstract network of "nodes" of space, in which each node was connected to every other. Very soon afterwards, this network collapsed and some of the nodes broke away from each other, forming the large universe we see today.

INFORMATION THEORY

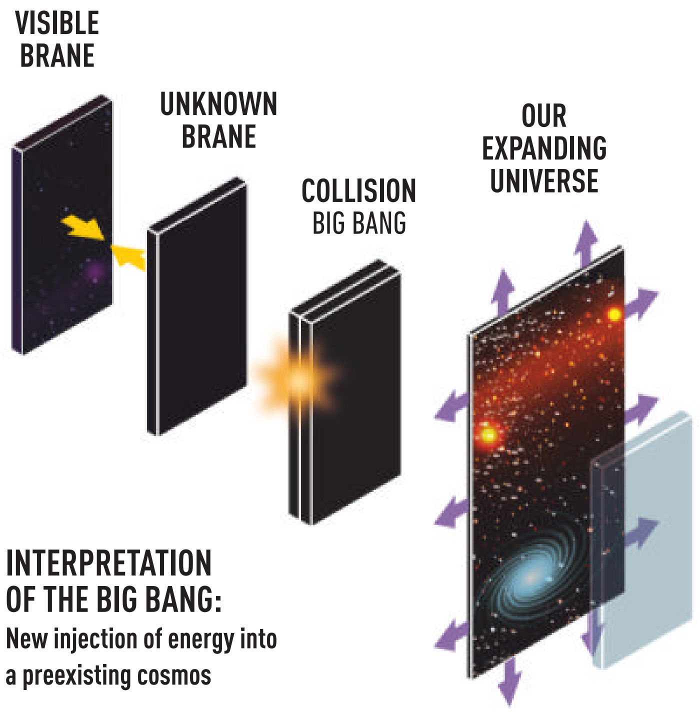
The usual focus on general relativity and quantum theory ignores the part that the third great pillar of physics, thermodynamics, might play. The laws of thermodynamics rule everything, including the flow of information. A better understanding of how they do this might provide clues as to why things are as they are – and narrow down the parameters of any ultimate theory.

BEYOND THE BIG BANG

What exactly happened at that point 13.8 billion years ago when our universe sprang into existence? This isn't something our current theories can explain fully convincingly. The inflationary big bang remains the most widely supported hypothesis, but when you add quantum theory into the mix, other possibilities become available.

COLLIDING BRANES

If our universe is a 4D "brane" floating in a higher-dimensional space, it might collide with other branes, acquiring energy that we interpret as the big bang



INTERPRETATION OF THE BIG BANG: New injection of energy into a preexisting cosmos

CYCLIC UNIVERSES

WAS the big bang perhaps not truly a beginning? That is one hypothesis cosmologists have put forward in recent years to explain away the inflationary big bang's problems. Rather than marking a "singularity" at the absolute beginning of space and time, it might just have been a recent event in a much longer history.

←
Turn back to page 30 for the basics on the inflationary big bang

The idea was first put forward in 2001 by Paul Steinhardt, who had become dismayed by the consequences of the inflationary theory he had helped develop two decades earlier, together with his colleagues Justin Khoury, Burt Ovrut and Neil Turok. The inspiration for this idea came from string theory, which predicts the existence of extra dimensions beyond the four of space and time we see.

This leads to the possibility that our 4D cosmos is situated on a "brane", a lower-dimensional object floating in a higher-dimensional space. In this picture, ours isn't necessarily the only brane: two branes floating a microscopic distance away from each other can form the boundaries to a five-dimensional space in between, like two slices of bread bounding a sandwich.

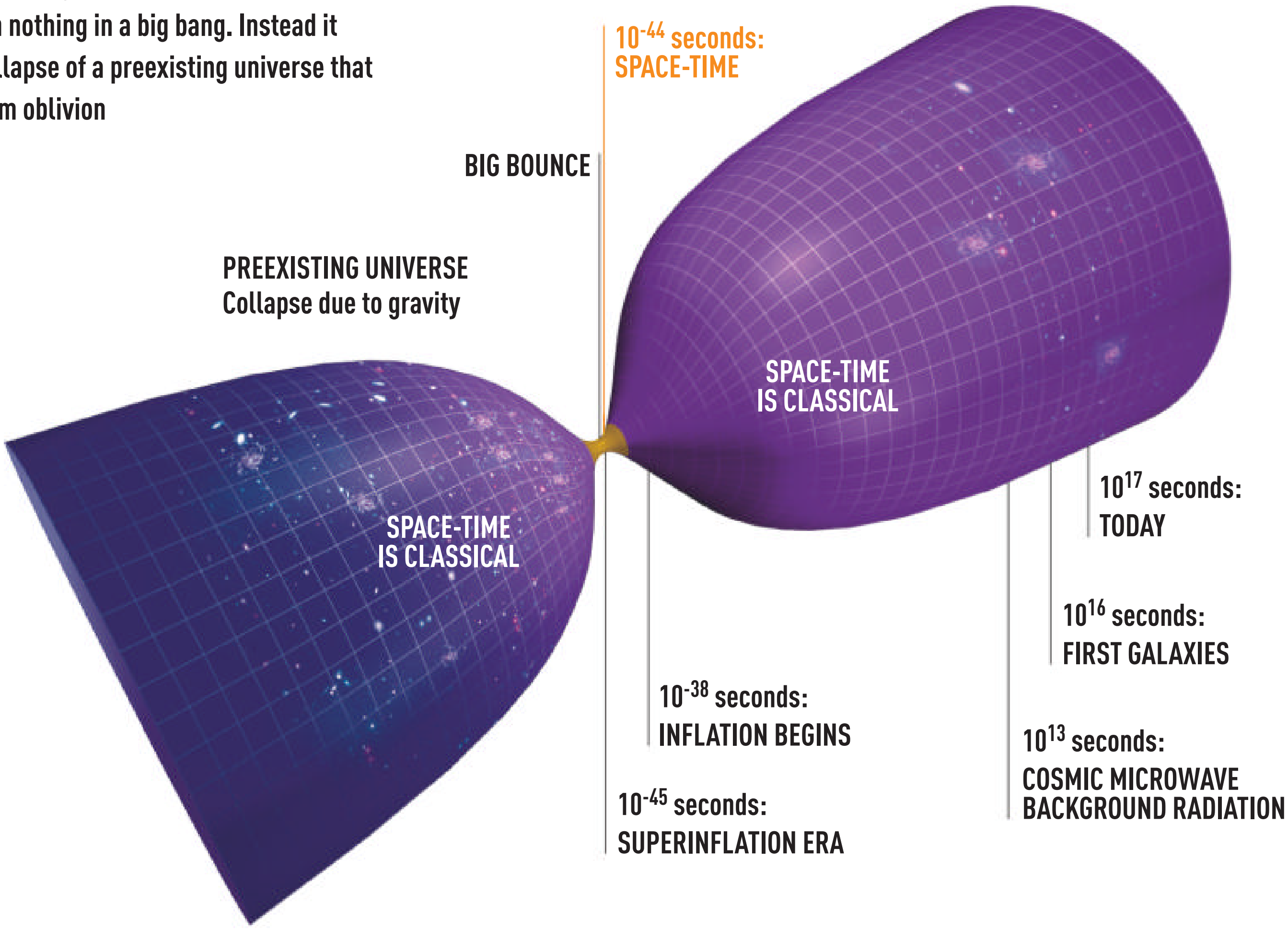
Steinhardt and his colleagues' idea was that every few trillion years or so, neighbouring branes float together and touch, resulting in an explosive exchange of energy as the fifth dimension briefly disappears into a singularity, reappearing as the branes move apart again (see diagram "Colliding branes", left).

Our 4D brane receives a huge injection of energy in this collision, something we interpret as a big bang. It doesn't represent the absolute beginning of our space and time, however, but a reinvigoration of a cosmos with an eternal existence – a "big bounce", if you like.

Such "ekpyrotic" universes do much of what the inflationary big bang was invented to do. "You can solve problems that would have been intractable without inflation given only 14 billion years," says Steinhardt. The branes are essentially flat to begin with, so the flatness problem disappears. When the branes clash, they hit at almost the same time everywhere, so the energy that will form matter and radiation pours

BOUNCE BACK

Loop quantum cosmology predicts that the universe did not arise from nothing in a big bang. Instead it grew from the collapse of a preexisting universe that bounced back from oblivion



in almost uniformly, creating a nearly homogeneous cosmos. Small quantum energy fluctuations are all that are needed to give enough density variation to eventually seed galaxies. And because the idea needs no multiverse, the "measure problem" that makes making cosmological predictions impossible disappears with it.

Since then, many other cyclic universe models that propose some form of cosmos existed before the big bang have sprung up. If nothing else, they introduced some competition into the market. "It shows that you're not stuck with inflation – other ideas are possible," says Steinhardt.

Cyclic scenarios depend on the existence of a long phase of ultra-slow contraction before the bounce. Just as inflation required a special form of energy, the inflaton field, to drive rapid expansion, ultra-slow contraction requires a special form of energy that exerts extraordinarily high pressure.

The high pressure slows contraction by resisting compression and, at the same time, tends to smooth out any irregularities in the distribution of energy and in the fabric of space-time. Unlike an inflationary phase, however, a slowly contracting phase doesn't require special starting conditions. It can be triggered in various ways, for example by decaying dark energy. What's more, in a slowly contracting, cold universe, quantum fluctuations remain small at all times. That means the outcome of the bouncing scenario is definite, unlike in the messy multiverse scenario, which is produced by wild quantum fluctuations during inflation.

In 2018, cosmologist Anna Ijjas at the Max Planck Institute for Gravitational Physics in Potsdam, Germany, published the first theoretical account of how such a bounce could happen, producing the kind of smooth distribution of energy and flat, untwisted geometry of space-time that we observe.

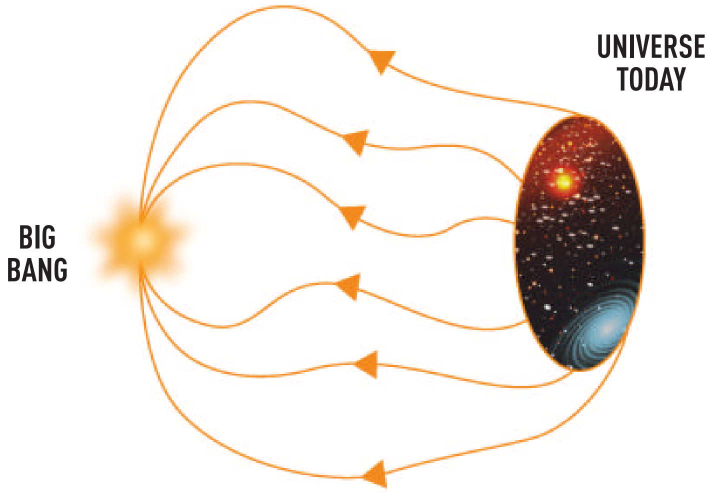
A natural extension of the concept is that we could be living in a cyclic universe with bounces occurring every 100 billion years or so. It is even possible to imagine a cyclic universe with no beginning or end. Each period of ultra-slow contraction would erase any fine details of the previous cycles and bring the universe to the bounce point with the same conditions as it had the cycle before.

As a result, all the features of the universe would be the same on average during each cycle, including the temperature, the concentration of dark matter, ordinary matter and dark energy and the number of observable stars and galaxies.

In other words, if you had lived on a planet like Earth in the cycle before our own, you would observe roughly the same basic properties of the universe as we do. ›

NO BOUNDARIES

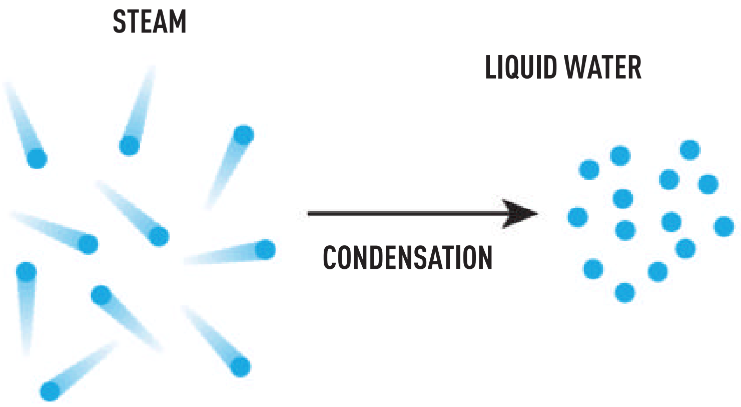
By adding the possible quantum histories of the universe together, we can work back from today's cosmos to the universe's origin



INTERPRETATION OF THE BIG BANG:
Timeless point at the beginning of many possible quantum histories where classical physics breaks down

CHANGING PHASE

Did the big bang create space and time, or could they have existed already in a bizarre, unfamiliar form? Water provides clues



When steam turns into liquid water, the molecules get closer together and are less able to move randomly

Others have produced cyclic models from different starting points, for example loop quantum gravity, a rival to string theory for the title of a viable theory of everything. If "loop quantum cosmology" turns out to be right, our universe emerged from a pre-existing universe that had been expanding before contracting due to gravity. As all the matter squeezed into a microscopic volume, this universe approached the so-called Planck density, 5.1×10^{96} kilograms per cubic metre. At this stage, it stopped contracting and rebounded, giving us our universe (see diagram "Bounce back", previous page).

That is because an extraordinary repulsive force develops in the fabric of space-time at densities equivalent to compressing a trillion solar masses down to the size of a proton. At this point, the quanta of space-time can't be squeezed any further. The compressed space-time reacts by exerting an outward force strong enough to repulse gravity. This momentary act of repulsion causes the universe to rebound. From then on, the universe keeps expanding because of the inertia of the big bounce. Nothing can slow it down – except gravity.

Such models lead to a dramatic prediction: the current phase of the universe in which its expansion rate is slowly accelerating will come to an end and the universe will enter a new contracting phase. It will then head towards a new bounce and a new phase of expansion. If that is true, the dark energy that is driving the current accelerated expansion must decay away, which may be detectable in future experiments.

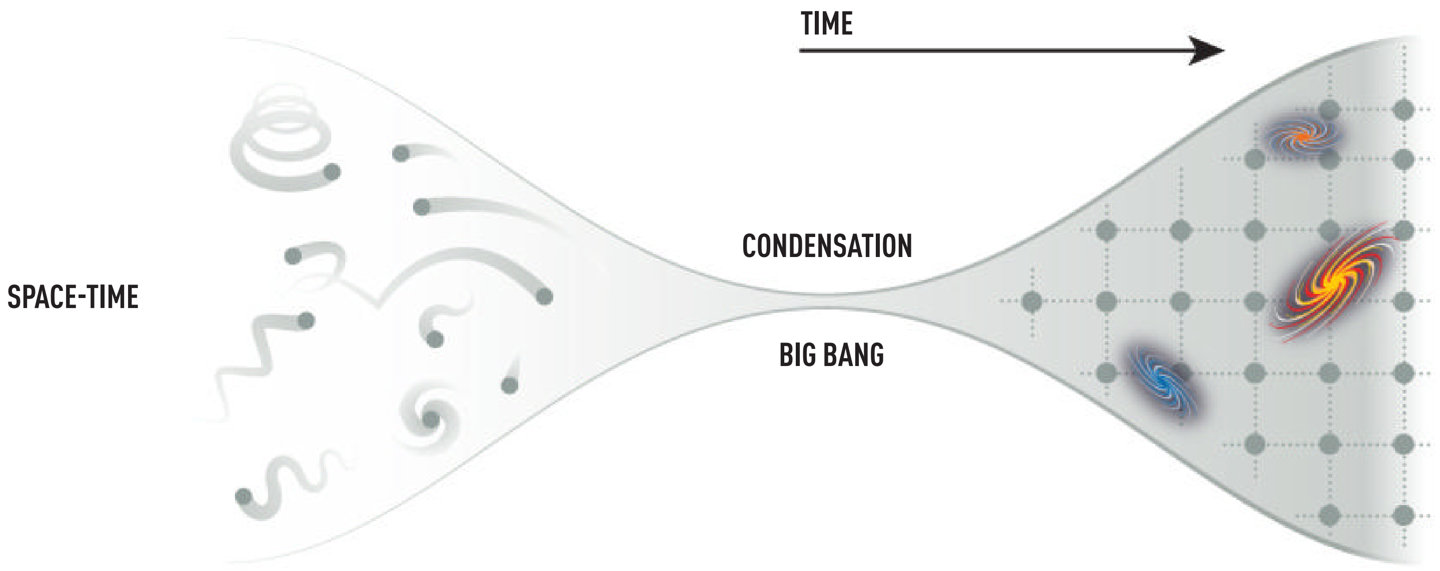
NO-BOUNDARY UNIVERSES

Models of the big bang that involve a singularity in our space-time, including the inflationary big bang, neatly excuse us from explaining what happened at the universe's beginning: the singularity is a place where the universe falls off the cliff of existence and the laws of physics break down. A criticism of cyclic models is that they don't explain how extra dimensions and pre-existing universes survive their momentary lapse into a singularity. "To me, it doesn't seem to work," says Thomas Hertog at the Catholic University of Leuven (KU Leuven) in Belgium, who worked on the idea for a couple of years. "The calculations suggest that the transition through the singularity is very unlikely."

The many clashes between branes that the model implies just compound the problem, says Sean Carroll at the California Institute of Technology. "If you follow the cyclic universe backward in time, the conditions that you need become more and more special, or unlikely." The way to solve that problem is to postulate some kind of beginning that provided a special set of conditions – but that seems to defeat the object of the theory.

Carroll thinks Steinhardt might have had the right idea in seeking an answer to the big-bang conundrum in the unification of general relativity and quantum mechanics, he says. But the answer could lie in using those ideas not to replace inflation, but to make it better.

"We use a half-assed version of quantum mechanics when we do cosmology," he says. Inflation is a theory about space-time and gravity, so it is anchored in general relativity. It incorporates a few aspects of



Space-time may also be built of atom-like parts. The big bang could have been when they "condensed", allowing familiar features like geometry and time to emerge

quantum physics, such as the uncertainty fluctuations that push the inflaton off the mountain ledge, but as we lack a sure-fire way of connecting relativity and quantum theory, it remains a "semiclassical" theory. "Maybe that's not good enough," says Carroll.

But how can we put the quantum into cosmology? Just after inflation had burst on to the scene in the 1980s, the late Stephen Hawking and his collaborator James Hartle at the University of California, Santa Barbara, made a stab at it. In quantum physics, when a particle travels from A to B, it doesn't take a single path but can pass along two or more paths simultaneously, interfering with itself at the other end as if it were a wave. To find out which path we are most likely to observe, we must add together the quantum-mechanical "wave functions" encoding each possible path, working out how their individual peaks and troughs cancel and amplify. Encoded within this total wave function is everything we need to know about the quantum particle at B, including the probabilities for the outcomes of any measurement we choose to make.

Hawking and Hartle argued that a similar approach could be applied to the universe as a whole. Point B is the universe we see today. Looking back towards its origin, we can trace many valid histories of its expansion back towards a point – point A – where semiclassical physics breaks down and quantum space and time become so gnarled that the two cease to be clearly distinguished. This point is no longer a beginning of time in a singularity as in the standard inflationary cosmology, but a timeless point where the universe – or rather a superposition of all possible historical universes – pops into existence from nothing with all its laws of physics intact. Because of this lack of a clear beginning, Hawking and Hartle called it the no-boundary proposal.

Following the rules of quantum mechanics, they added up all the possible histories that began in a universe with no boundary and ended in the universe we see today (see diagram "No boundaries", far top left). The idea is a kind of multiverse in reverse: a single universe with multiple histories. The resulting wave function gets rid of inflation's measure problem, as it encodes a unique set of probabilities for anything we might observe. And because the admissibility of the histories is determined by what we see in the cosmos today, problems such as the flatness of space-time or the homogeneity of the cosmic microwave background cease to be problems: instead, they are the inputs to the theory.

The no-boundary universe had its attractions for many physicists. "If it turns out the universe has to have an origin, I find this initial state to be a very plausible one," says Alan Guth, the originator of inflation. And as far as his theory is concerned, the no-boundary proposal turned up a pleasant surprise. As Hartle, Hawking and Hertog showed in 2008, although the theory wasn't dreamed up with inflation in mind, it crops up naturally along many paths the universe could have taken to get here. "You can calculate the probability that inflation occurred, and it turns out that probability is very high," says Hertog.

The idea also fits neatly with one of the most profound ideas to come out of string theory in recent years: the holographic principle. This states that the

›

physics of a 4D universe such as ours, including gravity, is mathematically equivalent to the physics on its 3D boundary without gravity. The implication is that the world we see around us is nothing but a holographic projection of information from the edge of reality. It sounds implausible, but the principle pops up not just in string theory, but in almost any approach to unifying relativity and quantum theory dreamed up so far.

Although the no-boundary proposal says that the universe has no boundary in the far past, it does give a boundary in the infinitely far future. By calculating the physics on this boundary, Hertog extracted the probabilities of all the possible universes that can emerge as its holographic projections. Remarkably, the probabilities for things like the homogeneity of the cosmic background or the amount of dark energy are the same as those that you get from the no-boundary wave function. This supplies a direct connection between string theory, the most popular route towards a theory of everything, and the no-boundary proposal, which produces inflation naturally. Of course, that is far from a concrete proof that it is a true picture of how the universe began – or didn't.

THE BIG BOIL

Imagine tipping a bucket of water over your head. The water is made of molecules that are ultimately governed by quantum theory, but you needn't be aware of the details to know you will be soaked. You could even work out precisely how the water would cascade using the science of hydrodynamics, which existed long before quantum theory. If hydrodynamics allows us to describe fluids without fussing over the fine details of molecules, might it be possible to create space-time from atoms of space, without first perfecting a description of those atoms?

That was an idea that came to Bei Lok Hu, a theorist at the University of Maryland in Baltimore, as a way of getting round the problem that no one knows what happens to space-time when, as we wind the story of the expanding universe backwards, the entire universe shrinks to a scale where quantum effects can't be ignored.

There is more to Hu's analogy than meets the eye. In the past few years, physicists have made models of warped regions of space-time from fluids, and found that the two are eerily similar. Think of water in its three familiar phases: ice, liquid water and steam. All are made of water molecules, but how those molecules interact varies. In steam, they whizz around, doing their own thing. If they hit a cold window pane, however, they begin huddling together, condensing from gas to liquid. Hu thinks space-time can undergo similar phase changes. Without something like condensation, the atoms of space would exist as some nebulous netherworld bereft of time and geometry.

Subsequent theoretical work has shown how such atoms could indeed undergo an orderly phase change into something resembling space, with basic features we take for granted, like geometry (see diagram "Changing phase", page 90). In 2017, further calculations showed that what emerged looked like the expanding space-time of our universe. And there was a surprise: the space-time fluid didn't like being funnelled into a singularity at the moment of the big bang. Instead, it wanted to bounce back outwards.

But Daniele Oriti, a theorist at the Max Planck Institute for Gravitational Physics in Potsdam, Germany, who was involved in that work, believes that once these models are examined more closely, there is every chance that the big event at the start of their universe will be neither a bang nor a bounce.

In such extreme conditions, he says, space-time could well have changed from one phase to another, meaning it didn't have a definite beginning at all. What we think of as the big bang was just the moment of condensation. The big condensation, you might say. Or, if we are still in rewind, the big boil.

So what is the space-time netherworld on the other side of the big boil like? Here, language fails, because every question – what, where, how – presupposes concepts that simply wouldn't have existed. "You have to think about these atoms of space without them existing somewhere in space, or evolving somewhere in time," says Oriti. "The very notion of time and space has to be constructed out of them." ∎

UNPICKING THE FABRIC OF THE UNIVERSE

Space-time is the fundamental fabric of the universe according to general relativity – but it might have a deeper origin in quantum theory, says Sean Carroll.

LET'S say you want to meet a friend for coffee. You have to tell them where you are going to be – your location in space – but you also need to let them know when. Both bits of information are necessary because we live in a four-dimensional continuum: three-dimensional space and everything within it, from steaming coffee machines to stars exploding in faraway galaxies, all happening at different moments of one-dimensional time.

"Space-time" is simply the physical universe inside which we and everything else exists. And yet, even after millennia living in it, we still don't know what space-time actually is. Physicists have strived to work it out for more than a century. In recent years, many of us have been trying to figure out what might be the threads from which the fabric of reality is woven. We have ideas, each with its own selling points and shortcomings. But for my money, the most exciting one is the most surprising.

It is the idea that space-time emerges from a weird property of the quantum world that means particles and fields, those fundamental constituents of nature, can be connected even if they are at opposite ends of the universe. If that is correct, we might finally have found a bridge between the two irreconcilable totems of physics, placing us on the threshold of a theory of quantum gravity. We would also have the most startling demonstration yet that the world we see isn't the world as it is – that there is always "something deeply hidden", as Albert Einstein put it – and that the only way to understand the fundamental nature of reality is by confronting quantum mechanics head on.

Space-time is a relatively new notion. Isaac Newton had no need for it. For him, space and time were individually real and absolute. Only when Einstein formulated his special theory of relativity in 1905 did the two start to come together. He showed that different observers will generally divide space-time into "space" and "time" in different, incompatible ways; what is "space" and what is "time" are relative to how an observer is moving.

›

PROFILE
SEAN CARROLL

Sean Carroll is a physicist at the California Institute of Technology, and author of books including *Something Deeply Hidden: Quantum worlds and the emergence of spacetime*

Various thinkers had previously speculated that the two should be rolled together. In Edgar Allan Poe's 1848 prose poem *Eureka*, for instance, he wrote that "space and duration are one". But it wasn't until 1908 that mathematician Hermann Minkowski unified them in a scientific way. He dramatically proclaimed: "Henceforth, space for itself, and time for itself, shall completely reduce to a mere shadow, and only some sort of union of the two shall preserve independence."

Einstein was unimpressed, grumbling about "superfluous learnedness". But he eventually came round to the idea, putting the geometry of space-time firmly at the heart of his general relativity. It said that space-time isn't merely a static background in which things happen. It is a dynamic entity, warping and stretching under the influence of mass and energy. The curvature of space-time manifests itself to us as the force of gravity.

Still, it would seem weird to ask what space-time is "made of" in classical physics. In general relativity, space-time changes over time in response to other stuff. But it is still a background, and a fundamental constituent of nature. It isn't made of anything; it just "is".

The problems with that view started with the discovery of quantum mechanics, the rules that govern the behaviour of subatomic particles and fields. Scientists haven't been able to construct a quantum-mechanical theory of gravity as they have for the three other fundamental forces of nature. Part of the issue is technical: when we try to make classical general relativity into a quantum-mechanical theory using standard techniques, our equations blow up and we get nonsensical answers. But part of it is conceptual.

Quantum mechanics tells us that systems exist in superpositions of different measurable quantities like position and velocity. There is no such thing as "the position" of a quantum particle; there are many possible positions, which take on definite values only when we observe them. How in the world can space-time exist in a superposition of different possibilities? That would make it impossible to say for sure that a certain event happened at a definite location in space and time.

Physicists of different persuasions have taken different approaches to constructing a solution in the form of a theory of quantum gravity. The most popular is string theory, which replaces particles with loops and segments of vibrating string. String theory successfully produces a quantum version of gravity, but not one that connects with our world in an obvious way. Nor does it resolve those fundamental conceptual problems. String theory's leading rival, loop quantum gravity, is an attempt to directly quantise general relativity. Loop proponents typically take the conceptual challenges of quantum gravity more seriously than their stringy colleagues, but the challenges remain.

This has led some physicists to take a step back and ask the question in a different way. The standard approach to developing a quantum description of some phenomenon, like the electromagnetic field or a collection of atoms, is to start with a classical description and then "quantise" it. That approach has failed again and again when it comes to gravity and space-time. It also isn't how nature works. The real world doesn't start classically and then somehow quantise. It is quantum from the start, and the classical world emerges as an approximation.

So maybe we shouldn't be trying to quantise gravity at all. Perhaps we should instead formulate a quantum theory from the start, and then show how classical space-time emerges from that. It is a new approach that has dramatic consequences for how we think about what space-time itself is made of.

To make progress in this direction, it is helpful to start with our current best physical theory, which is quantum field theory. According to this theory, the fundamental ingredients of the world are fields, such as the electric and magnetic fields. Even particles like electrons and quarks are simply vibrations in fields that stretch through space.

Classically, we can specify the value of a field in an approximate fashion by dividing space into tiny regions, then listing the value of the field in each region. Once we graduate to quantum field theory, an extra feature comes into the game: the values of the field in different regions can be entangled with each other. Due to quantum uncertainty, we don't know exactly what answer we will get if we measure the field at some location, but entanglement means the answer we get at one point will affect what we would measure at any other point.

In the vacuum state of an ordinary quantum field theory – empty space, no particles flying around – the entanglement between fields in different regions is directly tied to the distance between them, and therefore to the geometry of space-time. Nearby regions are highly entangled with each other, while faraway regions share little entanglement.

This suggests an intriguing way to reverse our normal way of thinking and so find space-time within quantum theory. Let us imagine starting with just a quantum state, no preexisting notion of space-time. Now we can try to work backwards, to extract space-time from entanglement.

If in ordinary physics the entanglement between two regions goes down as the regions get further apart, let us imagine defining the distance as (inversely) related to the entanglement. In that case, having a quantum state automatically gives us the "distance" between any two parts of it, and therefore defines a geometry on this emergent space.

So far, so good. But a quantum state exists at each moment of time, so it can at best define the geometry of space at that moment. We want to extend this to four-dimensional space-time. Thankfully, here we can borrow a trick from physicist Ted Jacobson at the University of Maryland in Baltimore, who, in 1995, showed how we could derive Einstein's equation for general relativity from simple assumptions about the relationship between entropy and geometry.

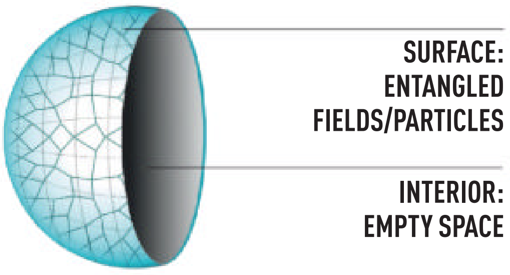
Entropy, a measure of disorder, is directly related to entanglement: the more entangled a region is with the rest of the world, the more entropy it contains. Einstein said that it is adding matter or energy to a region that causes space-time to curve. Jacobson showed that increasing the entanglement of a region can have the same effect, if we insist that the amount of entropy must be proportional to the area bounding that region. That is automatically true in empty space, but Jacobson suggested that it remains true even when space isn't empty. You can try to add more entanglement, but space-time will bend to compensate, so that entropy always remains proportional to area.

So Einstein says that energy causes curvature, while Jacobson says entanglement does. But Jacobson also argued that it is really the same thing: whenever you add entanglement, energy necessarily follows. From this logic, he was able to derive that the curvature of space-time in his approach obeyed the same equation that Einstein first wrote down for general relativity. Gravity, it appears, can arise from entanglement, rather than directly from mass and energy. This remarkable result was the beginning of what is now called "thermodynamic" or "entropic" gravity.

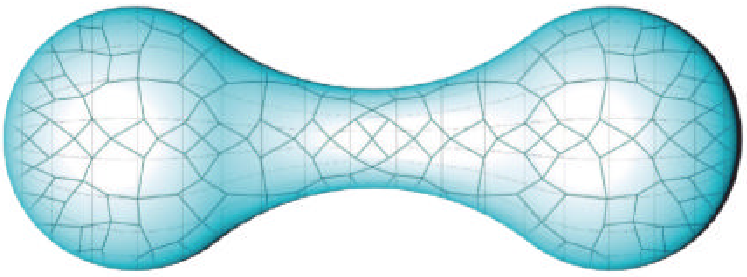
But it doesn't quite get us to where we need to be. ›

Space-time might be woven from quantum entanglement

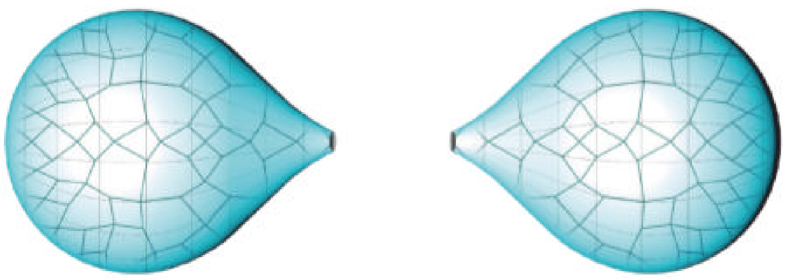
In a model universe, the equations describing gravity in a volume of space are equivalent to those describing the surface of that volume, which don't include gravity. This suggests space on the inside somehow emerges from the properties of the outside, namely entanglement



Sure enough, when you reduce the entanglement connecting two regions of the outside surface, the space inside pulls apart as if pulling at two ends of a piece of chewing gum



If you eliminate all entanglement, the space inside splits in two, suggesting that entanglement is the thread that binds space-time



In deriving his alternative picture of where gravity comes from, Jacobson assumed a classical space-time and imagined that there were quantum fields living within it. Ideally, we would like to keep everything quantum from the start and derive the existence of space-time itself. This is something I recently attempted with my collaborators, ChunJun (Charles) Cao and Spiros Michalakis at the California Institute of Technology. Rather than starting with vibrating quantum fields living in space-time, we started with completely abstract quantum "degrees of freedom".

This is just some quantity that can take on different values, independently of other quantities. In Newtonian mechanics, the degrees of freedom are positions and velocities of particles; in field theory, they are the values and rates of change of the fields. In our approach, the degrees of freedom don't have any direct physical interpretation. They are the basic stuff of reality, the essence out of which everything else is made – a kind of "quantumness" that preexists everything. Most importantly, these quantum degrees of freedom are entangled with each other.

With that in mind, we flip around Jacobson's idea. Now we can define the area surrounding a region as the entanglement of its degrees of freedom with the outside world. And sure enough, the corresponding geometry obeys Einstein's equation of general relativity. Gravity, in other words, can emerge directly from the quantum essence of reality, without quantising any assumed classical stuff.

That might sound like a conclusion, but it is more like a promising beginning. Many assumptions went into our derivation, and whether these assumptions hold true in nature remains to be seen. Most importantly, our derivation of Einstein's equation from entanglement only works when gravity is weak and space-time is nearly flat. Once gravity becomes strong and space-time is curved, as in the big bang or near a black hole, radically new phenomena become important.

The most dramatic of these is the holographic principle, the idea that the degrees of freedom describing a black hole can be thought of as living on its edge, the event horizon, rather than the interior. Juan Maldacena at the Institute for Advanced Study in Princeton, New Jersey, used the holographic principle to show an equivalence between two very different theories: quantum field theory without gravity in four-dimensional space-time, and quantum gravity with a negative vacuum energy in five dimensions.

Subsequent work by Mark Van Raamsdonk at the University of British Columbia in Canada and others has shown that the space-time geometry on the quantum-gravity side of this correspondence is directly tied to quantum entanglement on the field-theory side. As we decrease entanglement in the field theory, space-time on the quantum-gravity side grows apart (see diagram, previous page).

Maldacena and Leonard Susskind at Stanford University in California have taken this connection to extremes with a bold idea they dubbed "ER = EPR." ER stands for Albert Einstein and Nathan Rosen, who wrote a paper in 1935 proposing the existence of wormholes, or shortcuts through space-time. EPR, meanwhile, stands for Einstein, Boris Podolsky and Rosen, who collaborated on another paper emphasising the role of entanglement in quantum theory. The ER = EPR conjecture therefore posits that whenever you have two entangled particles, there is a tiny wormhole connecting them.

Don't take this too literally. The wormholes that purportedly connect pairs of particles would be microscopically small and impossible for anything to pass through. It is only when massive amounts of entanglement become involved that we begin to see a macroscopic distortion in the fabric of space.

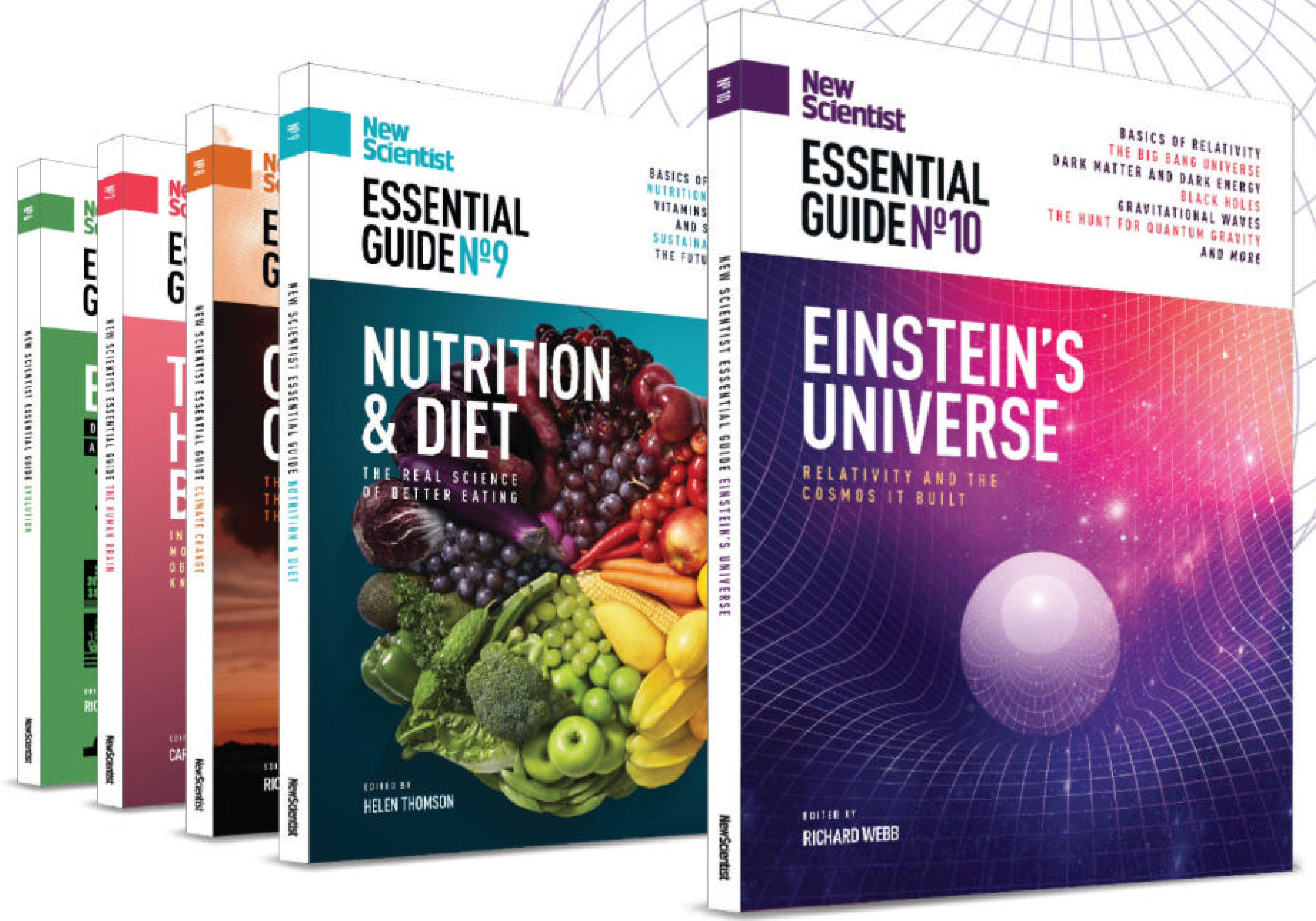
Moreover, our universe has a positive vacuum energy, not a negative one, so the implications of the equivalence revealed in Maldacena's negative-vacuum-energy thought experiment don't translate directly to an actionable strategy for dealing with quantum gravity in the real world. They do, however, serve as another strong hint that quantum entanglement is at the heart of it all.

All of these ideas are, at present, somewhere between promising conjectures and optimistic dreams. We don't know the best way to think about these supposed fundamental degrees of freedom that entangle together to make space-time, nor do we know how they interact with each other in any detailed way. We can't yet derive the emergence of quantum fields living within space-time, obeying the rules of relativity. And we certainly can't answer important questions like why the energy of empty space is so small, or why space has four macroscopic dimensions.

Even so, imagining that space-time emerges from quantum entanglement is a promising way to think about the basic nature of reality. It may be that it was a mistake to start with general relativity and try to quantise it; maybe space-time was lurking within quantum mechanics all along.

And even if formulating a complete theory of quantum gravity isn't your thing, thinking about space-time this way should at least put a new slant on the familiar four-dimensional continuum in which we live, rushing around in space to be on time for coffee. ▮

NewScientist
ESSENTIAL GUIDES


New Scientist
ESSENTIAL GUIDE №9
BASICS OF
NUTRITION
VITAMINS
AND S
SUSTAINA
THE FUTU
NUTRITION & DIET
THE REAL SCIENCE OF BETTER EATING
EDITED BY
HELEN THOMSON
New Scientist
ESSENTIAL GUIDE №10
BASICS OF RELATIVITY
THE BIG BANG UNIVERSE
DARK MATTER AND DARK ENERGY
BLACK HOLES
GRAVITATIONAL WAVES
THE HUNT FOR QUANTUM GRAVITY
AND MORE
EINSTEIN'S UNIVERSE
RELATIVITY AND THE COSMOS IT BUILT
EDITED BY
RICHARD WEBB
NEW SCIENTIST ESSENTIAL GUIDE EINSTEIN'S UNIVERSE
NEW SCIENTIST ESSENTIAL GUIDE NUTRITION & DIET
NEW SCIENTIST ESSENTIAL GUIDE CLIMATE CHANGE
NEW SCIENTIST ESSENTIAL GUIDE THE HUMAN BRAIN
NEW SCIENTIST ESSENTIAL GUIDE EVOLUTION

NEW SCIENTIST ESSENTIAL GUIDES

DELIVERED DIRECT TO YOUR DOOR

Based on the best coverage from *New Scientist*, the *Essential Guides* are comprehensive, need-to-know compendiums covering the most exciting themes in science and technology today.

Get the series, including the recently published issue on nutrition & diet, with an *Essential Guides* subscription. It means you don't have to search for issues in the shops – we can deliver them direct to your door.

FOR MORE INFORMATION ON FUTURE ISSUES AND SUBSCRIPTION OFFERS, VISIT:

NEWSCIENTIST.COM/ESSENTIALGUIDE

New Scientist

ESSENTIAL GUIDE Nº10

EINSTEIN'S UNIVERSE

HOW DID THE UNIVERSE BEGIN, AND HOW IS IT CHANGING? WHAT IS IT MADE OF? DO BLACK HOLES TRULY EXIST?

ALBERT EINSTEIN'S SPACE-AND-TIME-WARPING THEORIES OF RELATIVITY HAVE REVOLUTIONISED OUR VIEW OF THE COSMOS OVER THE PAST CENTURY. FIND OUT HOW IN THIS TENTH *NEW SCIENTIST* ESSENTIAL GUIDE, WITH TOPICS INCLUDING:

① How relativity works

② The big bang universe

③ Dark matter and dark energy

④ Black holes and gravitational waves

⑤ The hunt for quantum gravity

£9.99

9 772634 015019 10